

SINUSOIDAL MODEL ANALYSIS-BY-
SYNTHESIS FOR SPEECH
CODING

By

WALTER D. ANDREWS

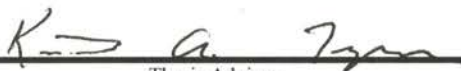
Bachelor of Science
Oklahoma State University
Stillwater, Oklahoma
1993

Master of Science
Oklahoma State University
Stillwater, Oklahoma
1994

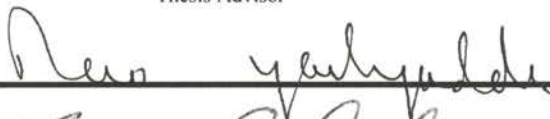
Submitted to the Faculty of the Graduate College of the
Oklahoma State University in partial fulfillment
of the requirements for the Degree of
DOCTOR OF PHILOSOPHY
May, 1998

SINUSOIDAL MODEL ANALYSIS-BY-
SYNTHESIS FOR SPEECH
CODING

Thesis Approved:



Thesis Advisor



George Scheets



Wayne B Powell

Dean of Graduate College

PREFACE

This study describes the application of analysis-by-synthesis for determining the parameters corresponding to the sinusoidal model for reconstruction. The error metric adopted is the mean-squared error. Two novel approaches for solving the minimum mean-squared error problem for the sinusoidal model are presented. These are referred to as frequency-domain analysis-by-synthesis and time-domain analysis-by-synthesis.

This would not have been possible without help. There are number of people I would like to thank. First, I would like to thank the Department of Defense for providing funding for my work over the last five years.

I would like to thank the members of my advisory committee, Dr. Keith Teague, Dr. Rao Yarlagadda, Dr. George Scheets, and Dr. Art Pentz. A very special thanks is extended to Dr. Keith Teague for his guidance and friendship during my tenure at Oklahoma State University. Also, I am very grateful to the School of Electrical and Computer Engineering for the employment, which made this study possible.

The next person I would like to acknowledge and dedicate this dissertation to is Ronda Andrews, my wife. She is as responsible for the work presented here as I. She provided constant motivation and love in my efforts to attain a Ph.D. in Electrical Engineering. When the times were tough she inspired me to continue forward. Without Ronda this work would never have been completed. Thank You.

Also a very special thanks goes out to my two sons, Chuck and Caleb. They both were very patient and understanding during the completion of this study as well as providing me with a tremendous amount of motivation to finish.

Special thanks to my Mother, Kay Parker, and Father, Walter Andrews Jr., for their continued support over the last eight years. Last I would like to express my appreciation to my Grandparents, Woodrow and Lorene Lawrence for their support and encouragement.

TABLE OF CONTENTS

Chapter	Page
1 INTRODUCTION	1
1.0 Introduction	1
1.1 Sinusoidal Model Analysis-By-Synthesis	2
1.2 Overview of Dissertation	2
1.3 Previous Work	4
1.4 Organization of Dissertation	4
2 OVERVIEW OF SPEECH CODING	6
2.0 Introduction	6
2.1 Speech Production	8
2.2 Speech Properties	10
2.3 Hearing	15
2.4 Waveform Coding	17
2.4.0 Introduction	17
2.4.1 Time-Domain Waveform Coding	18
2.4.1.1 Pulse Code Modulation	18
2.4.1.2 Differential Pulse Code Modulation	19
2.4.1.3 Delta Modulation	20
2.4.2 Frequency-Domain Waveform Coding	21
2.4.2.1 Filter Bank Coding	21
2.4.2.2 Subband Coding	21
2.4.2.3 Adaptive Transform Coding	22
2.4.3 Vector Quantization	22
2.5 Voice Coding	24
2.5.0 Introduction	24
2.5.1 Vocoder Overview	24
2.5.2 Code Excited Linear Prediction	26

Chapter		Page
	2.5.3 Sinusoidal Transform Coder	29
	2.5.4 MultiBand Excitation	33
2.6	<i>Summary</i>	42
3	ENHANCED MBE	43
3.0	<i>Introduction</i>	43
3.1	<i>Analyzer</i>	45
	3.1.0 Introduction	45
	3.1.1 Pre-filtering and Windowing	45
	3.1.2 Pitch Estimate	48
	3.1.3 Pitch Refinement	53
	3.1.4 Voicing	58
	3.1.5 Spectrum	61
	3.1.6 Gain	67
3.2	<i>Quantizer</i>	71
3.3	<i>Synthesizer</i>	74
	3.3.0 Introduction	74
	3.3.1 Spectral Filtering	76
	3.3.2 Unvoiced Synthesis	80
	3.3.3 Voiced Synthesis	83
	3.3.4 Reconstructed Output	86
3.4	<i>Results and Conclusion</i>	87
4	SINUSOIDAL MODEL	90
4.0	<i>Introduction</i>	90
4.1	<i>Sinusoidal Model</i>	91
4.2	<i>Conclusion</i>	102
5	SINUSOIDAL MODEL ANALYSIS-BY-SYNTHESIS	103
5.0	<i>Introduction</i>	103
5.1	<i>Overview</i>	105
5.2	<i>Analysis</i>	109
5.3	<i>Synthesis</i>	114

Chapter		Page
5.4	<i>Frequency-Domain Analysis-By-Synthesis Sinusoidal Model</i>	121
5.4.0	Introduction	121
5.4.1	Frequency-Domain Analysis-By-Synthesis	122
5.4.2	Simulation	132
5.4.2.0	Introduction	132
5.4.2.1	Test Signals	134
5.4.2.2	Sub-Sampling Process	136
5.4.2.3	Analysis-By-Synthesis Loop	147
5.4.2.4	Match Scores	156
5.4.2.5	Sub-Sampling Period Contour	161
5.4.2.6	Synthesized Test Signals	164
5.4.2.7	Conclusion	167
5.5	<i>Time-Domain Analysis-By-Synthesis Sinusoidal Model</i>	168
5.5.0	Introduction	168
5.5.1	Time-Domain Analysis-By-Synthesis	170
5.5.2	Simulation	180
5.5.2.0	Introduction	180
5.5.2.1	Test Signals	182
5.5.2.2	Sub-Sampling Process	182
5.5.2.3	Analysis-By-Synthesis Loop	182
5.5.2.4	Match Scores	191
5.5.2.5	Sub-Sampling Period Contour	195
5.5.2.6	Synthesized Test Signals	198
5.5.2.7	Conclusion	201
6	SINUSOIDAL MODEL ANALYSIS-BY-SYNTHESIS VOCODER	203
6.0	<i>Introduction</i>	203
6.1	<i>Analyzer</i>	204
6.1.0	Introduction	204
6.1.1	Pre-Filtering and Windowing	206

Chapter		Page
	6.1.2 Pitch Estimate	208
	6.1.3 Voicing	208
	6.1.4 Spectrum	210
	6.1.5 Gain	211
6.2	<i>Quantizer</i>	212
6.3	<i>Synthesizer</i>	213
	6.3.0 Introduction	213
	6.3.1 Spectral Filtering	215
	6.3.2 Synthesis	216
	6.3.3 Reconstructed Output	217
6.4	<i>Conclusion</i>	217
7	CONCLUSION	219
	7.0 <i>Introduction</i>	219
	7.1 <i>Frequency-Domain Analysis-By-Synthesis</i>	220
	7.2 <i>Time-Domain Analysis-By-Synthesis</i>	222
	7.3 <i>Sinusoidal Model Analysis-By-Synthesis Vocoder</i>	223
	7.4 <i>Future Research</i>	224
	7.4.0 Introduction	224
	7.4.1 Frequency-Domain Analysis-By-Synthesis	224
	7.4.2 Time-Domain Analysis-By-Synthesis	226
	BIBLIOGRAPHY	227
	APPENDIX	230
A1	<i>Synthetic Phase</i>	230
A2	<i>Postfilter</i>	233
	Vita	237

LIST OF FIGURES

Figure		Page
Figure 2-1.	Vocal Tract	9
Figure 2-2.	Voiced Speech Signal	12
Figure 2-3.	Spectrum of Voiced Speech Signal	13
Figure 2-4.	Unvoiced Speech Signal	14
Figure 2-5.	Spectrum of Unvoiced Speech Signal	14
Figure 2-6.	Peripheral Auditory System	15
Figure 2-7.	Block Diagram of the Traditional Speech Production Model	25
Figure 2-8.	CELP Analyzer	27
Figure 2-9.	Frame of Voiced Speech with Amplitude and Frequency Estimates	30
Figure 2-10.	Frame of Unvoiced Speech with Amplitude and Frequency Estimates	30
Figure 2-11.	Original Spectrum for a Frame of Speech, $S_w(\omega)$	35
Figure 2-12.	Spectral Envelope for Original Spectrum, $H_w(\omega)$	36
Figure 2-13.	All Voiced Synthetic Spectrum, $P_w(\omega)$	37
Figure 2-14.	All Unvoiced Synthetic Spectrum, $U_w(\omega)$	38
Figure 2-15.	Estimated Voicing Decisions	38
Figure 2-16.	Mixed Excitation Spectrum, $E_w(\omega)$	39
Figure 2-17.	Estimated Synthetic Spectrum, $\tilde{S}_w(\omega)$	39
Figure 3-1.	Block Diagram of EMBE Analyzer	44
Figure 3-2.	Block Diagram of EMBE Synthesizer	44

Figure		Page
Figure 3-3.	Frequency Response for Input HPF	46
Figure 3-4.	Block Diagram for Initial Pitch Estimate	49
Figure 3-5.	Frequency Response of Residual LPF	50
Figure 3-6.	Block Diagram for Pitch Refinement	54
Figure 3-7.	Block Diagram for Estimating Voicing Decisions	59
Figure 3-8.	Block Diagram of Spectral Modeling	62
Figure 3-9.	Block Diagram for Gain Calculation	68
Figure 3-10.	Block Diagram of Quantization and Coding	71
Figure 3-11.	Block Diagram for Reconstructed Speech	74
Figure 3-12.	Block Diagram for Spectral Filtering	77
Figure 3-13.	Block Diagram of Unvoiced Synthesis	81
Figure 3-14.	Block Diagram of Voiced Synthesis	83
Figure 3-15.	Block Diagram for Reconstructed Speech	86
Figure 3-16.	Overlapping Tapered Window	87
Figure 5-1.	Block Diagram of Analysis Procedure	109
Figure 5-2.	Block Diagram of Synthesis Procedure	115
Figure 5-3.	Reconstruction Using a Hamming Window	119
Figure 5-4.	Frequency-Domain Analysis-By-Synthesis Using Sinusoidal Model	121
Figure 5-5.	All Voiced Signal	135
Figure 5-6.	All Unvoiced Signal	135
Figure 5-7.	The Word “Figure”	136
Figure 5-8.	All Voiced Magnitude Spectrum Sub-Sampled with $P_s = 20$	137
Figure 5-9.	All Voiced Magnitude Spectrum Sub-Sampled with $P_s = 114$	138

Figure		Page
Figure 5-10.	All Voiced Magnitude Spectrum Sub-Sampled with $P_s = 62.8$	139
Figure 5-11.	All Voiced Magnitude Spectrum Sub-Sampled with $P_s = 63.2$	140
Figure 5-12.	All Voiced Magnitude Spectrum Sub-Sampled with $P_s = 63.4$	140
Figure 5-13.	All Unvoiced Magnitude Spectrum Sub-Sampled with $P_s = 20$	141
Figure 5-14.	All Unvoiced Magnitude Spectrum Sub-Sampled with $P_s = 114$	142
Figure 5-15.	All Unvoiced Magnitude Spectrum Sub-Sampled with $P_s = 75$	143
Figure 5-16.	Speech Magnitude Spectrum Sub-Sampled with $P_s = 20$	144
Figure 5-17.	Speech Magnitude Spectrum Sub-Sampled with $P_s = 114$	145
Figure 5-18.	Speech Magnitude Spectrum Sub-Sampled with $P_s = 61.2$	146
Figure 5-19.	Speech Magnitude Spectrum Sub-Sampled with $P_s = 62.4$	146
Figure 5-20.	Original Magnitude Spectrum and All Voiced Synthetic Magnitude Spectrum for $P_s = 20$	147
Figure 5-21.	Original Magnitude Spectrum and All Voiced Synthetic Magnitude Spectrum for $P_s = 114$	148
Figure 5-22.	Original Magnitude Spectrum and All Voiced Synthetic Magnitude Spectrum for $P_s = 62.8$	149
Figure 5-23.	Original Magnitude Spectrum and All Voiced Synthetic Magnitude Spectrum for $P_s = 63.2$	150
Figure 5-24.	Original Magnitude Spectrum and All Voiced Synthetic Magnitude Spectrum for $P_s = 63.4$	150
Figure 5-25.	Original Magnitude Spectrum and All Unvoiced Synthetic Magnitude Spectrum for $P_s = 20$	151
Figure 5-26.	Original Magnitude Spectrum and All Unvoiced Synthetic Magnitude Spectrum for $P_s = 114$	152
Figure 5-27.	Original Magnitude Spectrum and All Unvoiced Synthetic Magnitude Spectrum for $P_s = 75$	152
Figure 5-28.	Original Magnitude Spectrum and “Figure” Synthetic Magnitude Spectrum for $P_s = 20$	153
Figure 5-29.	Original Magnitude Spectrum and “Figure” Synthetic Magnitude Spectrum for $P_s = 114$	154

Figure		Page
Figure 5-30.	Original Magnitude Spectrum and “Figure” Synthetic Magnitude Spectrum for $P_s = 61.2$	155
Figure 5-31.	Original Magnitude Spectrum and “Figure” Synthetic Magnitude Spectrum for $P_s = 62.4$	155
Figure 5-32.	Match Scores for Integer Sub-Sampling Periods of the All Voiced Signal	157
Figure 5-33.	Match Scores for Fractional Sub-Sampling Periods of the All Voiced Signal	158
Figure 5-34.	Match Scores for Integer Sub-Sampling Periods of the All Unvoiced Signal	159
Figure 5-35.	Match Scores for Fractional Sub-Sampling Periods of the All Unvoiced Signal	159
Figure 5-36.	Match Scores for Integer Sub-Sampling Periods of the Word “Figure”	160
Figure 5-37.	Match Scores for Fractional Sub-Sampling Periods of the Word “Figure”	161
Figure 5-38.	All Voiced Sub-Sampling Period Contour	162
Figure 5-39.	All Unvoiced Sub-Sampling Period Contour	163
Figure 5-40.	“Figure” Sub-Sampling Period Contour	164
Figure 5-41.	Synthetic All Voiced Signal	166
Figure 5-42.	Synthetic All Unvoiced Signal	166
Figure 5-43.	Synthetic Word “Figure”	167
Figure 5-44.	Time-Domain Analysis-By-Synthesis Using Sinusoidal Model	169
Figure 5-45.	Original Input Signal and All Voiced Synthetic Signal for $P_s = 20$	183
Figure 5-46.	Original Input Signal and All Voiced Synthetic Signal for $P_s = 114$	184
Figure 5-47.	Original Input Signal and All Voiced Synthetic Signal for $P_s = 62.76$	185
Figure 5-48.	Original Input Signal and All Voiced Synthetic Signal for $P_s = 63.13$	185
Figure 5-49.	Original Input Signal and All Voiced Synthetic Signal for $P_s = 63.5$	186

Figure		Page
Figure 5-50.	Original Input Signal and All Unvoiced Synthetic Signal for $P_s = 20$	187
Figure 5-51.	Original Input Signal and All Unvoiced Synthetic Signal for $P_s = 114$	187
Figure 5-52.	Original Input Signal and All Unvoiced Synthetic Signal for $P_s = 74.92$	188
Figure 5-53.	Original Input Signal and “Figure” Synthetic Signal for $P_s = 20$	189
Figure 5-54.	Original Input Signal and “Figure” Synthetic Signal for $P_s = 114$	189
Figure 5-55.	Original Input Signal and “Figure” Synthetic Signal for $P_s = 62.4$	190
Figure 5-56.	Original Input Signal and “Figure” Synthetic Signal for $P_s = 62.76$	191
Figure 5-57.	Original Input Signal and “Figure” Synthetic Signal for $P_s = 63.12$	191
Figure 5-58.	Match Scores for Sub-Sampling Periods of the All Voiced Signal	193
Figure 5-59.	Match Scores for Sub-Sampling Periods of the All Unvoiced Signal	194
Figure 5-60.	Match Scores for Sub-Sampling Periods of the Word “Figure”	195
Figure 5-61.	All Voiced Sub-Sampling Period Contour	196
Figure 5-62.	All Unvoiced Sub-Sampling Period Contour	197
Figure 5-63.	“Figure” Sub-Sampling Period Contour	198
Figure 5-64.	Synthetic All Voiced Signal	199
Figure 5-65.	Synthetic All Unvoiced Signal	200
Figure 5-66.	Synthetic Word “Figure”	200
Figure 6-1.	Block Diagram of SMABS Analyzer	205
Figure 6-2.	Block Diagram of Quantization and Coding	212
Figure 6-3.	Block Diagram of SMABS Synthesizer	214

1 INTRODUCTION

1.0 Introduction

The current interest in speech coding is attributed to the demand for voice communication, the new generation of technology for cost-effective implementation of digital signal processing algorithms, the need to conserve bandwidth, and the need to conserve disk space in speech storage [1], [2]. Although not a new topic, low bit rate speech coding has become an important area for research in recent years. There are several reasons for this occurrence. These include advances in microprocessor technology, the sharp decrease in cost of computation and memory, the increased emphasis on providing high-quality communication services, and the improvement in speech models [1], [2]. The result of the recent surge in research has been the development of a number of good speech coding systems at bit rates of 4,800 bits per second (bps) and below. Many of these voice coders (vocoders) are sinusoidal based systems known as harmonic vocoders, such as, Sinusoidal Transform Coding (STC) [3], Improved MultiBand Excitation (IMBE) [4], [5], and Enhanced MultiBand Excitation (EMBE) [6], [7]. While these vocoders produce high quality and intelligible speech at low bit rates, they are highly dependent on accurate parameter estimation.

1.1 Sinusoidal Model Analysis-By-Synthesis

This dissertation examines the problem of coding narrowband speech with an emphasis on using analysis-by-synthesis in combination with a sinusoidal speech model. Currently, there are no harmonic vocoders employing the use of analysis-by-synthesis for complete parameter estimation and synthesis. Although analysis-by-synthesis is not a new approach to speech coding it has not been coupled with the sinusoidal model. The inclusion of the sinusoidal synthesis model into the analyzer assures that the model parameters being estimated are optimal in some sense. This new approach to parameter estimation for sinusoidal based vocoders is referred to as Sinusoidal Model Analysis-By-Synthesis (SMABS).

1.2 Overview of Dissertation

The motivation for the work presented in this dissertation stems from the fact that low bit rate harmonic vocoders are in general highly dependent on the estimation of the pitch (fundamental frequency). Generally, pitch estimation is performed in an open-loop fashion, a number of heuristic tests are needed to maintain a smooth transition from frame-to-frame. For these reasons, a new approach for estimating the pitch and other necessary parameters is desired. The approach to parameter estimation investigated in this dissertation is that of analysis-by-synthesis coupled with the sinusoidal synthesis model.

A number of linear prediction analysis-by-synthesis systems that minimize a perceptually weighted residual signal (error signal) are discussed in [1], [2], and [9]. In the process of minimizing the error signal a pitch estimate, commonly known as the pitch

delay, is required. The pitch is found using either an open-loop or closed-loop approach. The pitch estimate determined does not necessarily correspond to the exact pitch of the speaker, since the residual is minimized using a perceptual criterion. As a result, the pitch estimate found in this manner is not accurate for sinusoidal based vocoders. But the techniques of the linear prediction analysis-by-synthesis system prove to be useful in developing sinusoidal model analysis-by-synthesis techniques for obtaining parameter estimates for sinusoidal based vocoders as is presented in Chapter 5.

Both MBE and STC are presented as analysis-by-synthesis vocoders. The difference between these vocoders and the work presented in this dissertation is in the model assumptions. MBE uses analysis-by-synthesis to perform pitch refinement by synthesizing an all voiced magnitude spectrum. STC uses analysis-by-synthesis to develop an alternate approach to pitch estimation without actually performing the synthesis in the analyzer. Another difference between MBE, STC, and the work developed in this dissertation is that neither MBE nor STC actually include the synthesis technique in the analyzer. The work in this dissertation takes another approach in using analysis-by-synthesis. The sinusoidal model for reconstruction is included directly in the analyzer forming a more complete approach to analysis-by-synthesis.

Two novel approaches of incorporating analysis-by-synthesis in a sinusoidal speech model are described. The first development is a frequency domain approach that uses a no phase information assumption. The second development is a time-domain approach that assumes phase information is available. For both methods developed, the computational complexity is considered to be of secondary importance.

A more practical sinusoidal vocoder is developed that exploits the advantages of each method developed. This new SMABS vocoder targets a bit rate of approximately 8,000 bps. In order to determine the level of performance achieved by the proposed methods, the objective performance measures Diagnostic Acceptability Measure (DAM) and Diagnostic Rhyme Test (DRT) are considered.

1.3 Previous Work

The work in this dissertation is the extension of work performed by the author in collaboration with others [6], [7]. A 2,400 bit per second vocoder based on the MBE speech model was developed as part of the Department of Defense Digital Voice Processing Consortium (DDVPC). This work was funded by the Department of Defense as part of the development of a new Federal Standard at 2,400 bit per second to replace Federal Standard 1015 LPC10e. This work and the work by [3], [4], [5], and [8] are combined to form a new analysis-by-synthesis vocoder based on the sinusoidal model.

1.4 Organization of Dissertation

The thesis starts with a general introduction into the areas of speech coding, speech production, speech properties, and hearing. This is followed by a brief discussion of waveform coding and voice coding. Chapter 3 describes the development and implementation of the Enhanced MultiBand Excitation 2,400 bps vocoder (EMBE). EMBE is the vocoder developed for the DDVPC and constitutes much of the work, which leads to the developments in this dissertation. Chapter 4 is a mathematical development of the application of the sinusoidal speech model to speech signals. This provides the

supporting material for the new methods developed in this dissertation. Chapter 5 introduces the concept of using analysis-by-synthesis with a sinusoidal speech model. A discussion of the analysis and synthesis methods precedes the development of the frequency-domain and time-domain analysis-by-synthesis techniques, which form the main body of this work. Chapter 6 uses the knowledge of the analysis-by-synthesis techniques of Chapter 5 to develop a more realizable analysis-by-synthesis sinusoidal vocoder targeted towards 8,000 bps. Chapter 7 summarizes the results of the frequency-domain and time-domain analysis-by-synthesis techniques along with the 8,000 bps analysis-by-synthesis sinusoidal vocoder.

2 OVERVIEW OF SPEECH CODING

2.0 Introduction

In order to completely understand the speech models described and used throughout this dissertation it is essential to understand at least the basics of the speech production process, speech properties, and the auditory system. By understanding these components of the human communication system, it should be possible to exploit the system properties in order to improve the performance of the analysis and synthesis stages of a speech coding system.

The digital processing of speech requires a transformation from the continuous domain to the discrete domain. This is accomplished by sampling at the appropriate sampling frequency F_s and quantizing the amplitudes of the samples to a finite range of values. The speech signals used in this dissertation are assumed to be narrowband, 0 – 4 KHz (Nyquist frequency), with a sampling frequency F_s set to be twice the Nyquist frequency, $F_s = 8$ KHz. A speech coder is then used to process the speech signal in preparation for transmission, storage, or encryption.

Generally, speech coders are separated into two categories: waveform and voice coders [1], [2], [9]. Waveform coders are used to represent and reconstruct accurately a digital speech signal on a sample-by-sample basis. In this approach the shape of the

waveform is preserved. Efficient waveform coding is accomplished by two methods. The first method exploits the redundant properties of the speech signal and is known as time-domain waveform coding [10], [11], [12]. The second method, frequency-domain waveform coding, exploits the non-uniform distribution of speech information in the frequency spectrum [10], [11], [12]. The most limiting factor when using waveform coders is that the coded speech seriously degrades as bit rates drop below about 16 kbps [5].

In practical signal processing applications, analysis and synthesis is performed by splitting the signal into short time segments unless the signal of interest is of short duration. The short time segment is often referred to as a *frame*. Signals are divided into frames so that conventional analysis techniques can be applied to signals that exhibit nonstationary characteristics, such as speech [10]. It is well known that speech signals are considered stationary over the range of 10ms to 30ms.

The application of this dissertation is in the area of low bit rate coding, so a stronger emphasis is placed on voice coders. A voice coder is also often referred to as either a vocoder or a parametric coder. In contrast to waveform coding, vocoders do not attempt to preserve the shape of the waveform, rather the goal is simply to preserve the perceptual qualities of the speech signal such as naturalness and intelligibility. The most common parameters used to accomplish this are pitch, voicing, and vocal tract response. These parameters are quantized and transmitted to the receiver for processing. The motivation for vocoders is that they have shown to be more successful than waveform coding at producing high quality speech at bit rates below about 16 kbps [1], [2].

This chapter provides a brief discussion on the areas of speech production, speech properties, and hearing. The following section provides a discussion of the common techniques used in waveform coding. This chapter concludes with a discussion of the techniques used in voice coding with special emphasis on mixed excitation harmonic based models.

2.1 Speech Production

The human communication system consists of a transmitter and a receiver, where the transmitter is represented by speech production and the receiver is represented by hearing. In this section and sections following the components of the speech production, speech properties and hearing systems are presented.

The transmitter portion of the human communication system is speech production. The vocal organs that make-up the speech production system are the lungs, the windpipe, the larynx, the pharynx, the nose, and the mouth. These vocal organs form a tube that extends from the lungs to the lips. The lungs represent the energy source, the larynx represents the excitation source, and the other organs represent a resonating system of air filled cavities. All of these components are required to produce speech in the human communication system [13].

One component of this tube, which extends from the lungs to the lips, is known as the vocal tract. The vocal tract consists of the pharynx, mouth, and nose. The vocal tract is shown in Figure 2-1. The vocal tract starts at the vocal cords and ends at the mouth with a typical length, for an adult male, of 17 *cm*. The cross sectional area of the vocal

tract varies in size from 0 to 20 cm^2 and is determined by the positioning of the speech articulators, composed of the tongue, jaw, lips, and velum [12], [14].

The lungs produce a steady stream of air, which forms the energy source. This air stream is made audible using a number of methods but the most frequent method used is vocal cord action [41]. The vocal cords make-up an adjustable barrier across the air passage as pictured in Figure 2-1 [41]. When the vocal cords are open, air is passed into the vocal tract; when the vocal tract is closed, air is blocked from entering the vocal tract. During speech, the vocal cords open and close in a periodic fashion generating a series of puffs. This series of puffs generates a buzz whose frequency increases as the rate of vibration of the vocal cords is increased. The characteristics of speech are determined by the shape of the vocal tract, which is continually altered by movements of the tongue and lips. Speech generated using this method is classified as voiced sound.

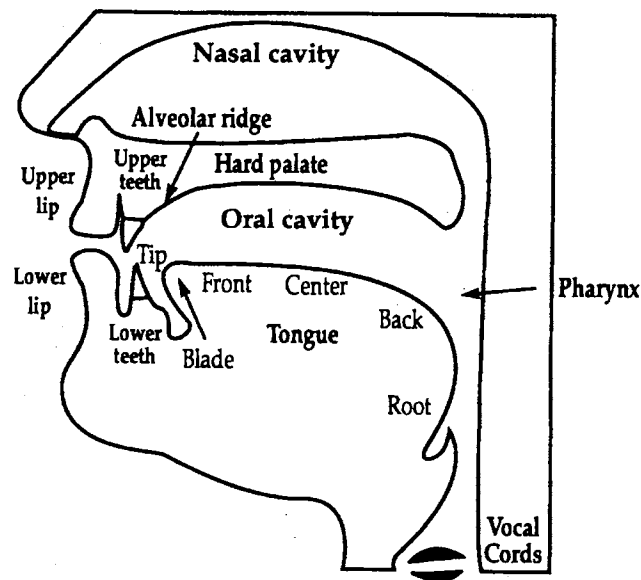


Figure 2-1. Vocal Tract [13]

There are two other methods that are used to produce speech waves. In the first, a constriction occurs at some point along the vocal tract causing the airflow from the lungs to become turbulent. This turbulent airflow produces a hiss sound, referred to as fricative noise. Speech generated in this manner is classified as unvoiced sound. The second method creates a momentary constriction blocking the airflow completely using the tongue or lips. The air pressure generated from this blockage is released suddenly resulting in sounds known as plosives. In both methods the vocal cords could vibrate simultaneously with the occurrence of the constriction [14], [41]. Fricatives and plosives can be classified as either voiced or unvoiced or mixed sound.

Another alternate path for sound occurs when the velum is lowered, acoustically coupling the nasal cavity with the vocal tract. When this occurs, the nasal sounds of speech are produced. The nasal tract is a non-uniform tube of fixed area and length with a typical length, for an adult male, of 12 *cm* [14].

2.2 Speech Properties

The vocal tract is an air-filled chamber that acts like a resonator. This means that the vocal tract responds more naturally to sound waves which are at the same frequency as the resonant frequencies of the vocal tract. If the vocal cords produce a series of puffs as stated in the previous section then the spectrum will contain a number of frequencies that occur at integer multiples of the fundamental frequency [41]. The fundamental frequency is the same as the vocal cords' frequency of vibration and corresponds to the

spectrum's lowest frequency component. As a sound wave such as this propagates along the vocal tract, the components corresponding to the resonant frequency are emphasized.

The vocal resonator has a number of resonant frequencies and will emphasize the harmonics of the vocal cord wave at a number of different frequencies [41]. These frequencies are determined by the shape of the vocal tract and change as the vocal tract shape is altered. These resonant frequencies are known as the formant frequencies. Each configuration of the vocal tract contains its own characteristic formant frequencies.

The formant frequency positions depend on the shape of the vocal tract, as noted above. If the soft palate is raised shutting off the nasal cavity the vocal tract is a tube approximately seven inches long [41]. Assuming that the vocal tract has uniform cross-section then the resonant frequencies occur at 500 Hz and its odd harmonics (1,500, 2,500, 3,500, etc.) [41]. In reality the cross-section of the vocal tract varies along its length, which results in a shifting of the frequency either higher or lower. The lowest formant frequency is known as the first formant, the next highest is the second formant, and so forth.

When the soft palate is lowered which results in a coupling of the nasal cavity and the mouth a different vocal tract shape is obtained [41]. This coupling provides two directions for air to be passed through the vocal tract. The addition of the nasal cavity introduces anti-resonances that suppress parts of the speech spectrum [41].

A voiced speech signal is shown in Figure 2-2 with its corresponding magnitude spectrum provided in Figure 2-3. This signal is a 240 point (30 ms) segment taken from the word "Figure" spoken by a male with a pitch of approximately 130 Hz. The

magnitude response is found by computing an 8,192 length DFT of the Hamming windowed speech segment.

The formant structure is clearly visible in Figure 2-3. The first formant occurs around 400 Hz, the second formant occurs around 1,800 Hz, the third formant occurs around 3,300 Hz and the fourth occurs around 3,700 Hz. A fact worth noting is that the voiced speech tends to be of high energy compared to unvoiced sounds, which is useful in determining whether the speech is either voiced or unvoiced.

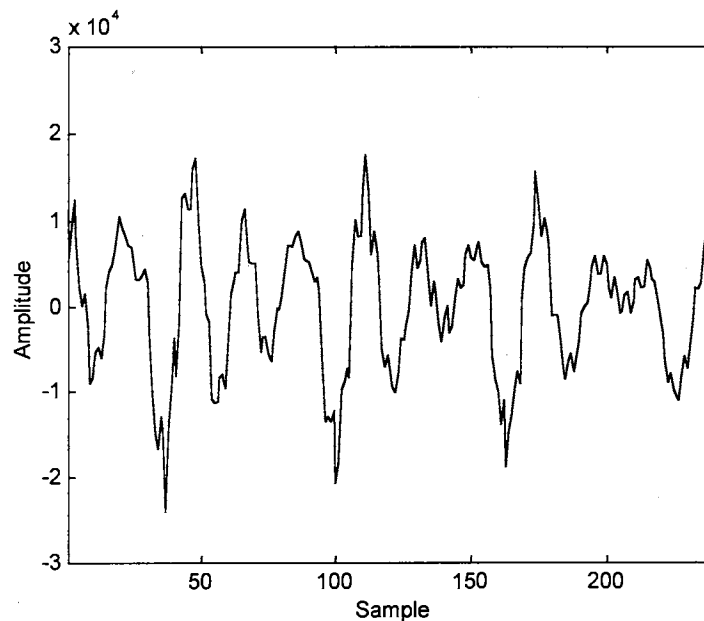


Figure 2-2. Voiced Speech Signal

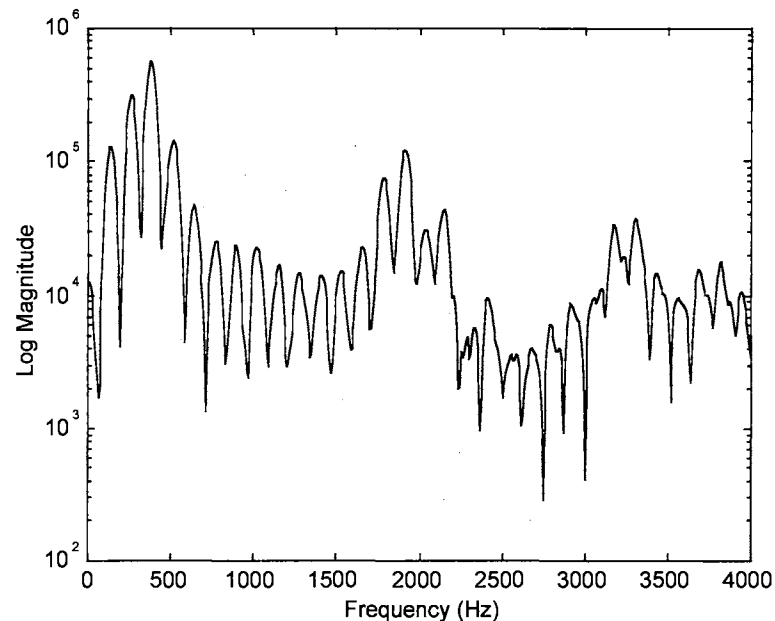


Figure 2-3. Spectrum of Voiced Speech Signal

Unvoiced sounds exhibit no harmonic structure since the vocal cords do not vibrate. This results in an aperiodic time-domain signal as shown in Figure 2-4. This aperiodic structure is clearly obvious in the corresponding magnitude spectrum provided in Figure 2-5. This speech segment is also a 240 point (30 ms) segment taken from the word “Figure” and spoken by a male. Again the magnitude response is found by computing an 8,192 length DFT of the Hamming windowed speech segment.

The formant structure that is so obvious for voiced speech is gone along with the harmonic structure. In contrast to voiced speech, unvoiced speech also tends to have lower energy. The unvoiced spectrum also appears to be a high pass signal in contrast to the low pass shape of the voiced spectrum.

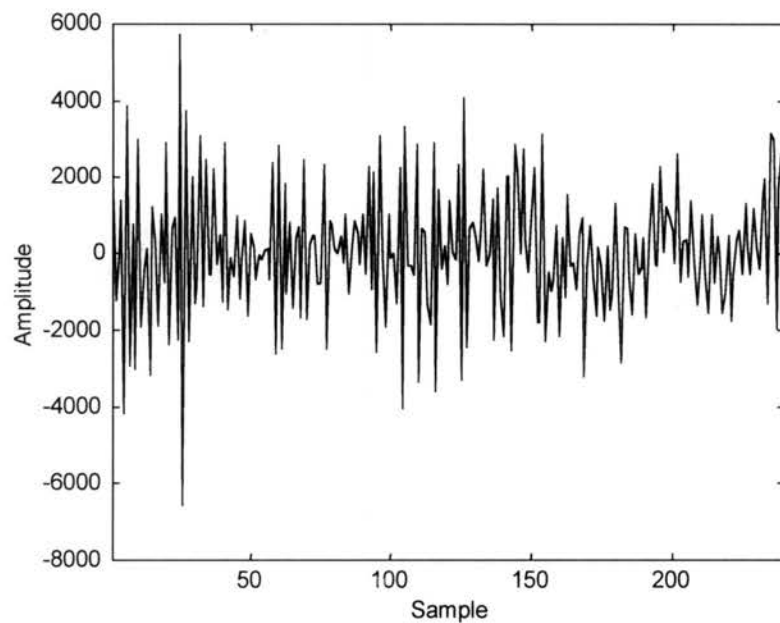


Figure 2-4. Unvoiced Speech Signal

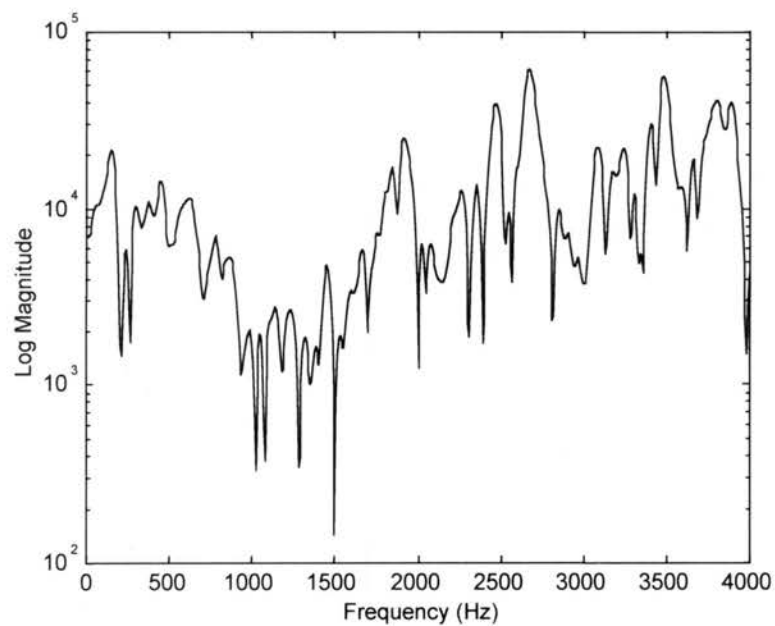


Figure 2-5. Spectrum of Unvoiced Speech Signal

2.3 Hearing

The speech production represents the transmitter portion of the human communication system. Hearing and perception represent the receiver portion. Of the two components the perceptual portion is the least understood in terms of how the brain decodes the acoustic information received. However, the detection of acoustic signals is fairly well understood [15]. A simple diagram illustrating the auditory system is shown in Figure 2-6. The function of the ear of the human communication system is to receive acoustic vibrations and convert them into signals suitable for transmission along the auditory nerve toward the brain [41].

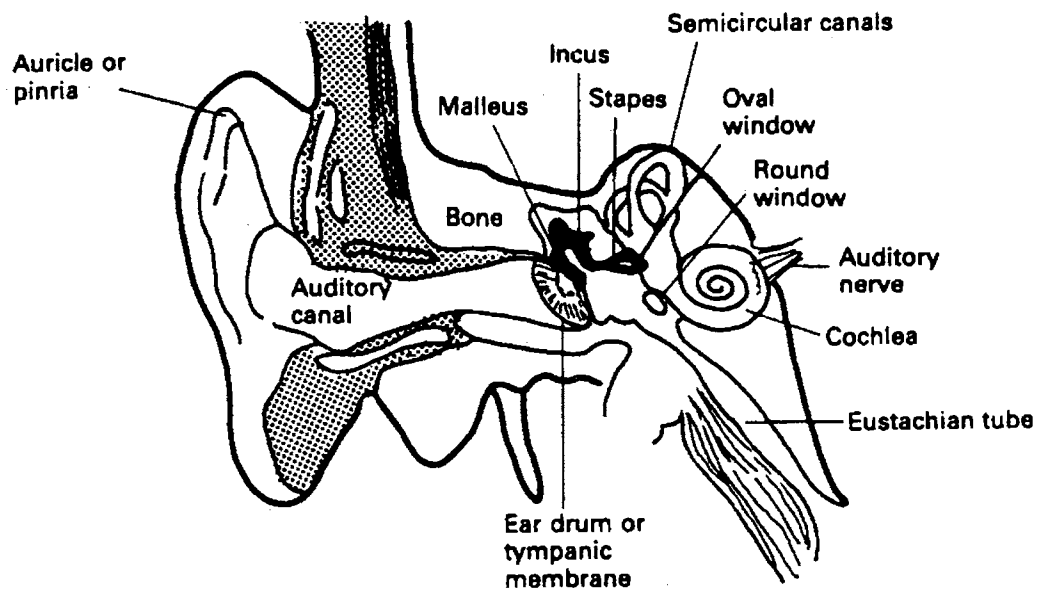


Figure 2-6. Peripheral Auditory System [14]

The makeup of the ear is generally divided into three main sections, outer ear, middle ear, and inner ear. The outer ear, which plays a minor role in the hearing process,

consists of the pinna, the auditory canal, and the eardrum [14], [15]. The auditory canal is closed at one end by the eardrum and open on the other end to the outside. As acoustic waves travel into the external ear, they are channeled down the auditory canal and subsequently set the eardrum into vibration. The auditory canal is an acoustic resonator so sound waves near the resonant frequency are amplified. The resonant frequency falls in the range of 3,000 Hz to 4,000 Hz [41].

The middle ear contains the auditory ossicles, three small bones that form a linkage between the eardrum and the inner ear [14], [15]. The three small bones are referred to as the malleus, the incus, and the stapes. The handle of the malleus (hammer) is connected to the eardrum [41]. As sound waves propagate down the auditory channel and strike the eardrum, the motions are transmitted to the stapes via the malleus and incus. The footplate of the stapes covers the oval window, which is the entrance to the inner ear [41]. The middle ear performs two functions. First it increases the amount of acoustic energy entering the fluid filled inner ear by increasing the amplitude of the pressure variations at the oval window. Without this function much of the incident energy would be reflected [41]. The second function is to protect the inner ear by suppressing violent vibrations. This is accomplished by opening the eustachian tube when a pressure difference is detected between the middle ear and the outer ear [14], [15].

The inner ear converts the vibrational energy into hydraulic energy and is subsequently converted to neurological signals for processing by the central auditory nervous system. This conversion of sound pressure waves occurs in the cochlea. The cochlea is a liquid filled tube with a helical (spiral) shape. Attached to the cochlea is a

frequency dependent membrane known as the basilar membrane. As pressure waves propagate through the cochlea, the basilar membrane is deflected at points corresponding to the frequency of the pressure waves. After the basilar membrane is deflected, a set of tiny hairs located at the organ of corti become bent creating a potential. The hairs are connected to the auditory nerve system and any potential causes a neural firing. These neural firings (series of electrical impulses) are transmitted to the brain for processing.

Based on the discussion above, one could think of the basilar membrane (or the ear in general) as a spectrum analyzer because of the frequency dependence. The ear is also known to act in a logarithmic manner with greater resolution at the lower frequencies [14], [15].

2.4 Waveform Coding

2.4.0 Introduction

Waveform coders are used to represent and reconstruct accurately a digital speech signal on a sample-by-sample basis, thus the shape of the waveform is preserved. Waveform coders are generally designed to be signal independent [15]. For this reason, waveform coders are considered to be extremely robust. This robustness comes at the expense of operating at a relatively high bit rate, such as 64 kbps.

The efficiency of the waveform coder is improved by exploiting the characteristics of the speech signal. In the time-domain the redundant property of the speech signal is exploited and is known as time-domain waveform coding [10], [11], [12]. In the frequency-domain the non-uniform distribution of speech information in the frequency

spectrum is exploited and is known as frequency-domain waveform coding [10], [11], [12]. The major problem with waveform coders is that the output speech seriously degrades as bit rates drop below about 16 kbps [5]. All the methods discussed in this section involve scalar quantization, except for the section on vector quantization.

2.4.1 Time-Domain Waveform Coding

2.4.1.1 Pulse Code Modulation

Pulse Code Modulation (PCM) is the most common method of time-domain waveform coding [10], [11], [12]. This method quantizes each sample of a speech signal to a specific discrete amplitude determined by the number of bits, B , used to represent the sample. The number of quantization levels is computed from B , as 2^B . The bit rate of a PCM coder is found by multiplying the number of bits, B , used to represent a given sample by the sampling frequency, F_s . Clearly, the more bits used for quantization the better the representation and the higher the bit rate.

Two types of scalar quantization are used with PCM, uniform and non-uniform. Uniform PCM has constant step size between quantization levels and non-uniform PCM has a step size that varies from quantization level to quantization level. The most common non-uniform PCM is the logarithmic quantizer. Two standard methods of logarithmic quantization are 8 bit A -law and μ -law PCM which are quite common in speech applications [9]. An optimum method for non-uniform PCM is to construct a quantization table based on the shape of the probability density function (pdf) for the typical speech data to be quantized.

A third approach is to make the quantization step size adaptive. This is accomplished using either a feedforward or feedback quantizer. A feedforward quantizer adapts the quantization step size based on the variance of a frame of speech. The step size must be transmitted as side information to the receiver in order to reconstruct the speech signal. The feedback quantizer adapts the quantization step size as a function of the previously quantized output. If the previously quantized output is small then a small step size is used. If the previously quantized output is large then the step size is increased accordingly. No side information needs to be transmitted to the receiver. Feedback adaptation produces lower bit rates but is more sensitive to transmission errors.

2.4.1.2 Differential Pulse Code Modulation

A variation on PCM is differential PCM (DPCM) [10], [11], [12]. The method of differential PCM takes advantage of the fact that samples close in time tend to be highly correlated. By exploiting this correlation the resulting bit rate is reduced. This is commonly accomplished by computing the difference between the current predicted sample and the adjacent sample. The difference between two adjacent samples has a lower dynamic range as compared to the original speech signal. Since the difference between two amplitudes is smaller on average than a particular amplitude, fewer bits are needed resulting in a lower overall bit rate. A common approach to DPCM is to use a linear predictor in the transmitter to estimate the current input sample from previous output samples. The difference between the original input and the estimate is quantized and transmitted along with the predictor coefficients. This method of waveform coding produces the same quality as PCM, but operating at a lower bit rate.

One possible method for improving the performance of DPCM is to allow the quantizer to adapt as is discussed in the section on PCM. This is known as Adaptive Differential Pulse Code Modulation (ADPCM). The quantizer is adapted to the prediction residual as are the prediction coefficients. Again, these terms are adapted using either a feedforward or feedback adaptation.

2.4.1.3 Delta Modulation

A simplified version of DPCM is known as delta modulation (DM) [10], [11], [12]. This system uses a sampling rate that is much higher than the Nyquist rate for the input signal. The result of the higher sampling rate is that adjacent samples become highly correlated.

The simplest DM coder uses a two-level quantizer (a '1' or '0') and a fixed first-order predictor to determine the quantized output signal with an associated quantization error [10], [11]. An equivalent realization for the fixed predictor is an accumulator with the input equal to the quantized error signal. The efficiency of the DM coder is constrained by two types of distortion, known as slope-overload and granular noise [10], [11]. Slope-overload distortion occurs when the quantization step size, the difference between two quantization levels, is too small to follow the steep slopes in the input waveform. Granular noise distortion occurs when the step size is too large to follow the small slopes in the input waveform.

For the reasons above a number of alternate approaches have been developed that use an adaptive step size to combat the distortion problem. One popular technique is known as continuously variable slope delta modulation (CVSDM) [10], [11]. This

approach uses a set of rules to adapt the step size. First a minimum and maximum step size are defined. If a run of three 1's or three 0's occurs then the step size is increased, if neither occurs the step decays until it reaches the minimum step size. This coder has been used mostly in areas where the speech quality does not have to meet commercial communications standards [10], [11].

2.4.2 Frequency-Domain Waveform Coding

2.4.2.1 Filter Bank Coding

The filter bank analyzer/synthesizer, which is usually considered a research coder, is representative of frequency-domain waveform coding [10], [11]. This waveform coder consists of a bank of bandpass filters that cover the entire frequency spectrum of interest. The speech signal is applied to the bandpass filters and the outputs are decimated for coding efficiency. The decimated outputs are quantized for transmission. In the synthesizer, the transmitted signal is interpolated and input to the same bank of bandpass filters. The outputs are then summed producing synthesized speech. This coder does not provide better coding efficiency, compared to the time-domain methods [10].

2.4.2.2 Subband Coding

An improvement on the filter bank analyzer/synthesizer is referred to as subband coding (SBC) [10], [11]. This method uses a bank of filters as in the previous method, but not as many. The frequency spectrum is divided non-uniformly into four to eight subbands, and each of these bands is encoded using either APCM or another waveform coding technique. This non-uniform frequency division is done because the low end of

the spectrum is considered perceptually more important than the high frequency end of the spectrum. More bits are assigned for coding the low frequency end of the spectrum. The high frequency end of the spectrum does not contain as much information so fewer bits are needed for coding. However, for this coder to achieve toll quality speech, four to five subbands and 24,000 bps is needed to code the entire spectrum.

2.4.2.3 Adaptive Transform Coding

Adaptive Transform Coding (ATC) segments the speech signal into frames of data, instead of filtering as in SBC [10], [11]. These frames are pushed into a buffer and then transformed into another form of representation, usually spectral. The transformed coefficients of the representation are quantized and transmitted to the synthesizer. At the synthesizer, the coefficients are inverse transformed back to the time domain. The bit rate is dependent on the number of bits used to code the coefficients. This type of coder has produced toll and communication quality speech at bit rates of 16 kbps and 9.6 kbps, respectively.

2.4.3 Vector Quantization

Vector quantization (VQ) is a generalization of scalar quantization techniques [16], [17], [18]. The main difference between scalar quantization and VQ is that scalar quantization operates on single samples while VQ performs operations on a set of ordered real numbers. A second difference is that scalar quantization is used primarily for analog to digital conversion and VQ is used in more sophisticated digital signal processing

applications where the signal has already been digitized [17]. VQ also exploits the linear and nonlinear dependence among signal vectors.

The method of VQ is the ultimate solution to the quantization of a signal vector [17]. The signal vector is compared to a set of similar but quantized signal vectors using a number of different error metrics, such as squared error, average squared error, or weighted squared error. The quantized signal vectors are stored into a table making the decoding process a simple task.

VQ is used primarily in the area of data compression, although it is not constrained to this application [16], [18]. In the area of speech coding, VQ has become quite popular as a method for quantizing the spectral envelope. The spectrum is first modeled using linear prediction (LP). The linear prediction coefficients are then converted to line spectral pairs, which are then coded using VQ. By using VQ, the line spectral pair coefficients are coded using two to three bits per coefficient [39]. This is compared to the INMARSAT standard, which uses 72 to 96 bits for coding the spectral envelope using scalar quantization [19]. Assuming a 14th order LP and 3 bits per coefficient then only 42 bits are needed to code the spectral envelope. This capability is exploited in Chapter 6 of this dissertation. The following section describes some of the common voice coders with a special emphasis on harmonic based coders.

2.5 Voice Coding

2.5.0 Introduction

The application of this dissertation is in the area of low bit rate coding, so a stronger emphasis is placed on voice coders. A voice coder is also known as either a vocoder or a parametric coder. In contrast to waveform coding, vocoders do not try to preserve the shape of the waveform but instead attempt to determine a set of time varying parameters that best describe the signal in terms of a speech production model. The set of parameters that is generally used to describe the speech production model are pitch, voicing, and vocal tract response. Vocoders have been shown to be more successful than waveform coding at producing high quality speech at bit rates of 16 kbps and below [1], [2].

2.5.1 Vocoder Overview

Vocoders differ from waveform coders because a mathematical model is used to represent the speech signal. This model estimates on a short time basis a set of parameters that is used to describe a frame of speech. The speech waveform is not necessarily preserved as in waveform coding, only the basic qualities are preserved. The model estimates are the inputs of a time varying linear system, as shown in Figure 2-7. The synthetic speech samples are represented by the output of the time varying linear system. The block diagram shown in Figure 2-7 is also a representation of what is referred to here as the *traditional* speech production model. The inputs: pitch, voiced/unvoiced decision,

gain, and vocal tract response represent the excitation parameters for each frame of output speech. A single voicing decision, modeled as a simple switch, determines whether the excitation is either periodic or aperiodic (voiced or unvoiced).

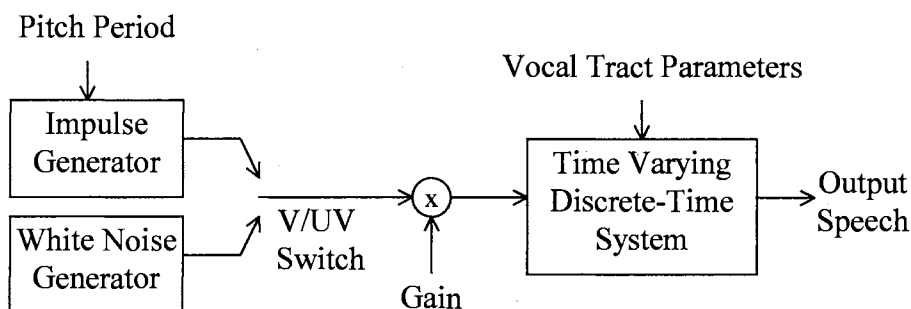


Figure 2-7 Block Diagram of the Traditional Speech Production Model

If the frame of speech is declared voiced, the excitation is modeled as a periodic impulse train with period equal to the pitch (commonly known as fundamental frequency). For a frame of speech declared unvoiced, the excitation is modeled as a pseudo-random white noise sequence. The vocal tract parameters are used to determine the spectral properties of the waveform for a frame of speech.

In all vocoders, a set of parameters must be estimated and updated periodically by the transmitter (analyzer). These parameters are usually the pitch, voiced/unvoiced decision(s), vocal tract response, and possibly an associated gain value. The parameters are encoded and transmitted to the receiver. In the synthesizer, the parameters are decoded once for every analysis frame and the speech signal is reconstructed on a frame-by-frame basis using the underlying speech production model.

The *traditional* speech production model described above and shown in Figure 2-7 has a major disadvantage. During analysis, a complete frame of speech is declared either periodic (voiced) or aperiodic (unvoiced). The approach of the simple voicing decision has been shown to be limiting because the resulting synthetic speech has an annoying “buzzy” quality. In modern speech coders, speech models have been based on the assumption that a given frame of speech is made up of both periodic and aperiodic excitation. Consider for example, voiced fricatives such as /v/ (“vice”), /D/ (“then”), /z/ (“zephyr”), and /Z/ (“measure”), which clearly contain mixed excitation [10]. This innovation has led to the development of a number of vocoders capable of producing high quality speech at low bit rates. The following paragraphs discuss in more detail a specific set of these vocoders.

2.5.2 Code Excited Linear Prediction

Code Excited Linear Prediction (CELP) was first introduced by Schroeder and Atal and is an analysis-by-synthesis method based on selecting the appropriate excitation sequence(s) as input to the synthesis filter [8]. The coefficients of the synthesis filter are found using linear prediction. A selected excitation sequence is input to the synthesis filter, which produces an estimate for the corresponding input speech signal. This estimate is subtracted from the original speech signal and the error is minimized using a weighted least square error approach. A simplified block diagram for the CELP analyzer is shown in Figure 2-8.

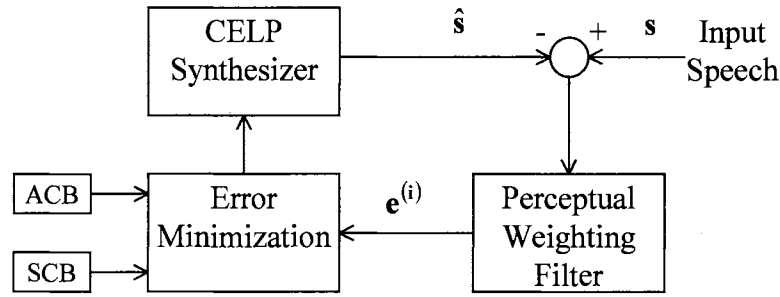


Figure 2-8. CELP Analyzer

There are two types of excitation sequences which are found by searching two codebooks. One codebook is used to model the periodicity in the error signal and is known as the Adaptive Codebook (ACB). The second codebook is used to model the randomness of the error signal, and is known as the Stochastic Codebook (SCB). These sequences are determined in a sequential manner [20], [21]. The ACB sequence which produces the minimal weighted least square error is chosen to represent the periodicity (voiced portion) for the corresponding input speech. The SCB sequence is added to the ACB sequence and the combination that produces the minimum weighted least square error is chosen to represent the unvoiced portion of the input signal. The minimum error is found by minimizing the following

$$\mathbf{e}^{(i)} = \mathbf{W}(\mathbf{s} - \hat{\mathbf{s}}^{(i)}). \quad (2-1)$$

The parameter \mathbf{W} is a matrix that represents the perceptual weighting shown in Figure 2-8. The perceptual weighting is used to flatten the spectrum of the error signal (residual). This is equivalent to weighting more equally the error over the entire frequency spectrum. An error value for each codebook index is computed and represented by the error vector $\mathbf{e}^{(i)}$. The reconstructed speech signal corresponding to each codebook

index is represented by the excitation vector $\hat{\mathbf{s}}^{(i)}$ and is determined by equation 2-2. The matrix \mathbf{H} represents the truncated impulse response of the synthesis filter, the vector \mathbf{u} is the contribution from the previous search stage, the vector $\mathbf{v}^{(i)}$ is the excitation sequence for the current search stage, and the vector $\hat{\mathbf{s}}^{(0)}$ is the zero input response of the synthesis filter (contribution from the previous frame).

$$\hat{\mathbf{s}}^{(i)} = \mathbf{H}(\mathbf{u} + \mathbf{v}^{(i)}) + \hat{\mathbf{s}}^{(0)} \quad (2-2)$$

CELP is one of the better-known high quality low bit rate vocoders. The CELP vocoder was adopted as Federal Standard 1016 at 4,800 bps in 1991 and is now widely used in a number of applications [21], [22], [23]. This vocoder is considered to be of good quality and intelligibility. CELP is often used as a reference when comparing the quality and intelligibility of other vocoders. For example, the goal of the new Federal Standard at 2,400 bps is to produce speech quality and intelligibility either equal to or better than that of the Federal Standard 1016 CELP 4,800 bps vocoder [21].

The strength of CELP lies in the analysis-by-synthesis method used to construct the excitation for the current frame. By feeding back the synthesized output that is produced by each possible excitation sequence chosen from the ACB and SCB, a feat made possible by including the synthesizer in the analyzer, an excitation sequence which is optimum in some sense is chosen. This provides for a degree of flexibility and a level of robustness that is not present in all vocoders. The concept of analysis-by-synthesis is explored later in this dissertation. The next section describes the Sinusoidal Transform Coder.

2.5.3 Sinusoidal Transform Coder

Sinusoidal Transform Coding (STC) is a speech analysis/synthesis system based on a sinusoidal model which was introduced by Robert J. McAulay and Thomas F. Quatieri in [3], [24]. This vocoder models the speech signal on a short time basis in terms of sums of sinusoids, where the sinusoids correspond in amplitude, frequency, and phase to the relative peaks in the spectrum of the current analysis frame. In practice, these peaks are estimated from the short-time Fourier transform (STFT). STC differs from the *traditional* speech production model because in the most general case there is neither a pitch estimate nor any voicing decisions. The receiver reconstructs the output speech by computing a sum of weighted sinusoids of corresponding frequencies and phases as determined in the analyzer. The frequencies are not in general related harmonically.

The analyzer uses a pitch adaptive Hamming window to segment the original speech, referred to as framing. The magnitude of the STFT is computed once for every frame. For each frame, the peaks (amplitudes) in the magnitude of the STFT are found by determining the frequencies where the amplitude slope changes from positive to negative. The peaks for a voiced and an unvoiced frame of speech are shown in Figures 2-9 and 2-10, respectively. The corresponding estimates for the frequencies $\hat{\omega}_l$ and phases $\hat{\theta}_l$ are then determined based on the location of the estimated peak \hat{A}_l . The variable l ranges from 1 to L , where L represents the number of peaks in the spectrum. These parameters are coded and transmitted to the receiver.

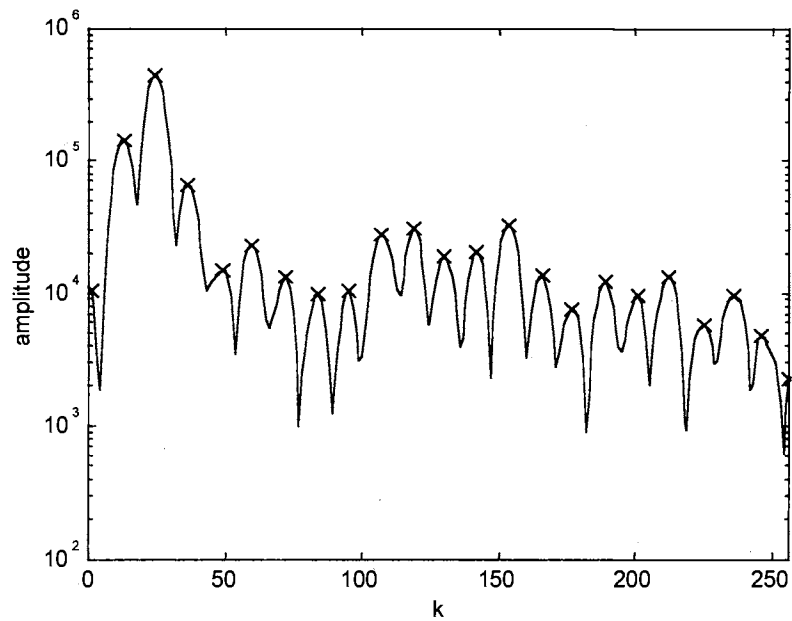


Figure 2-9. Frame of Voiced Speech with Amplitude and Frequency Estimates

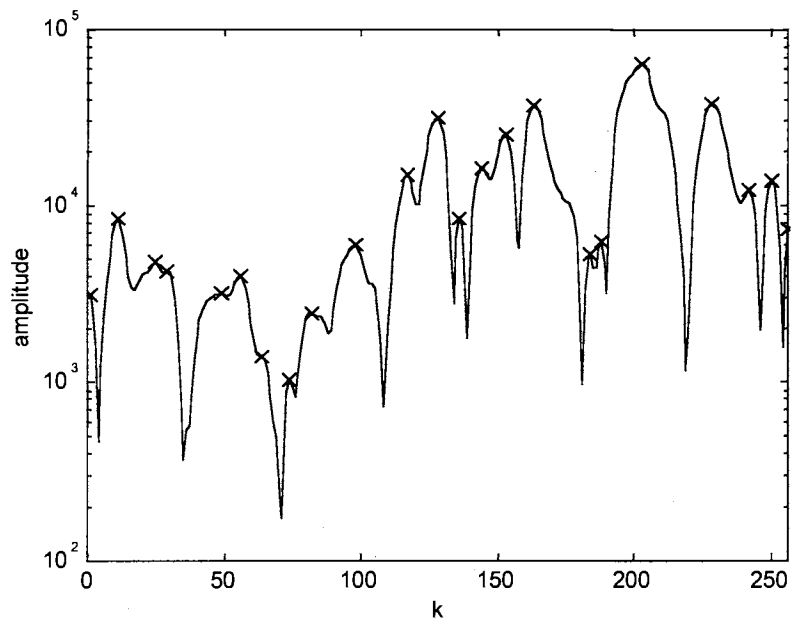


Figure 2-10. Frame of Unvoiced Speech with Amplitude and Frequency Estimates

The receiver decodes the transmitted parameters: amplitudes, frequencies, and phases. An amplitude modulated sinewave generator is used to generate a sinusoid for

each frequency that corresponds to the amplitude estimate and rotated by the corresponding phase estimate. These sinewaves are then summed to produce a frame of synthetic speech using

$$\tilde{s}[n] = \sum_{l=1}^L \hat{A}_l \cos(n\hat{\omega}_l + \hat{\theta}_l). \quad (2-3)$$

One major problem with this method is how to connect the sinewaves from frame to frame yet maintain smoothness across the frame boundaries. STC uses a cubic interpolation function to connect frequencies and phases smoothly from frame to frame as shown in equation 2-4 [3]. The parameters α and β are the unknown coefficients in the cubic polynomial and are found from equation 2-5.

$$\tilde{\theta}(t) = \hat{\theta} + \hat{\omega}t + \alpha t^2 + \beta t^3 \quad (2-4)$$

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \frac{3}{T^2} & \frac{-1}{T} \\ \frac{-2}{T^3} & \frac{1}{T^2} \end{bmatrix} \begin{bmatrix} \hat{\theta}_{+1} - \hat{\theta} - \hat{\omega}T + 2\pi M \\ \hat{\omega}_{+1} - \hat{\omega} \end{bmatrix} \quad (2-5)$$

The parameters in equations 2-4 and 2-5 are defined as follows: $\hat{\theta}$ is the phase estimate for the current frame, $\hat{\theta}_{+1}$ is the phase estimate for the future frame, $\hat{\omega}$ is the pitch estimate for the current frame, $\hat{\omega}_{+1}$ is the pitch estimate for the future frame, T is the size of the frame, and M determines the minimum number of cycles to track from the current frame to the next frame.

This sinusoidal model as presented would require a relatively high bit rate (10 Kbps or higher) because each peak in the spectrum must be transmitted along with its corresponding location (frequency) and phase. Since the direction of this dissertation is

towards the low bit rate vocoder, this model has to be slightly modified if it is be especially useful for low bit rate coding.

For voiced frames, the assumption is made that the sinusoids are harmonically related. With this assumption, only the pitch needs to be transmitted, as compared to sending every frequency location if the frequencies of the spectral peaks are not constrained. The location of the first sinusoid would represent the pitch, and subsequent sinusoids are located at harmonics of the pitch. It has been determined that an unvoiced spectrum sampled at approximately 100 Hz spacing (in frequency) is capable of reproducing noiselike signals using the sinusoidal model, thus the same sinusoidal model is used to represent speech which has voiced, unvoiced, or mixed-excitation [3]. The amplitudes are coded either directly, or using Linear Prediction (LP) or another alternate parametric model.

A parametric representation of the spectral amplitudes would result in a significant reduction in the number of bits necessary to code the amplitude information. The spectrum, if modeled using linear prediction, is encoded using vector quantization. The pitch is quantized using either 8 or 9 bits once per frame compared to the 8 or 9 bits per frame needed to encode the frequency location of each peak in the spectrum otherwise.

STC is claimed to produce high quality speech for various types of signals such as quiet speech, multispeaker waveforms, music, speech with background noise, and marine biological signals [3], [24]. A major advantage of STC is its ability to represent arbitrary

waveforms when sinusoids are unconstrained. Unfortunately this advantage vanishes if the sinusoids are constrained to have frequency related to a pitch and its harmonics.

The STC vocoder was a candidate for the Federal Standard at 4,800 bps, but finished well behind the other candidates in all tests [23]. STC was also a candidate for the new 2,400 bps Federal Standard developed by the Department of Defense Digital Voice Processing Consortium (DDVPC). While the STC type vocoder was not selected for standardization, it finished near the top.

The next vocoder described is MultiBand Excitation. MBE might be considered a special case of STC. The MBE model assumes that voiced frames have peaks in the spectrum that occur at the pitch and integer multiples. MBE is discussed in more detail in the following section.

2.5.4 MultiBand Excitation

MultiBand Excitation (MBE) is the speech model of most interest in this dissertation. This speech model was introduced in 1988 by D.W. Griffin and J.S. Lim [4], [5], [25], [26]. The *traditional* speech production model, presented earlier in this section, estimates three parameters to represent a frame of speech: pitch, a single voiced/unvoiced decision, and vocal tract response. The MBE speech model estimates the same three parameters but differs by assuming that both voiced and unvoiced excitation exist in the same frame (i.e., mixed excitation). Mixed excitation in a frame of speech is represented by splitting the spectrum into a set of predefined bands. Then each frequency band is declared either voiced or unvoiced. Splitting the spectrum into bands leads to multiple

voiced/unvoiced (V/UV) decisions per frame, as opposed to the *traditional* method of specifying only one voicing decision per frame. Of the various models for speech production which are under study, MBE is believed by many to hold the greatest promise for producing toll quality speech at low bit rates [1].

Since speech signals $s[n]$ are in general non-stationary, the speech data must be framed with a low pass analysis window $w[n]$ to focus the analysis on a short time interval. The time domain windowed speech segment is defined below as

$$s_w[n] = s[n]w[n]. \quad (2-6)$$

For MBE, mixed excitation is modeled in the frequency domain rather than the time domain as in CELP. The spectrum of the original windowed speech signal $|S_w(\omega)|$ is split into non-overlapping bands, and each band is modeled as being either voiced or unvoiced. Shown in Figure 2-11 is the original magnitude spectrum for a frame of speech. As is seen, the spectrum is voiced at the low end, then becomes unvoiced and then voiced again clearly displaying the existence of mixed excitation. The magnitude spectrum of the original speech signal $|S_w(\omega)|$ is modeled in the frequency domain as the product of a spectral envelope magnitude $|H_w(\omega)|$ and a mixed excitation magnitude spectrum $|E_w(\omega)|$ as shown in equation 2-7. This produces the synthetic magnitude spectrum $|\tilde{S}_w(\omega)|$. The vocal tract response is estimated by sampling the magnitude of the DFT spectrum at integer multiples of the estimated pitch. The spectral envelope magnitude

$|H_w(\omega)|$ for this frame of speech is provided in Figure 2-12. The vocal tract response is equivalent to the smooth spectral envelope of the original frame of speech.

$$|\tilde{S}_w(\omega)| = |H_w(\omega)| |E_w(\omega)| \quad (2-7)$$

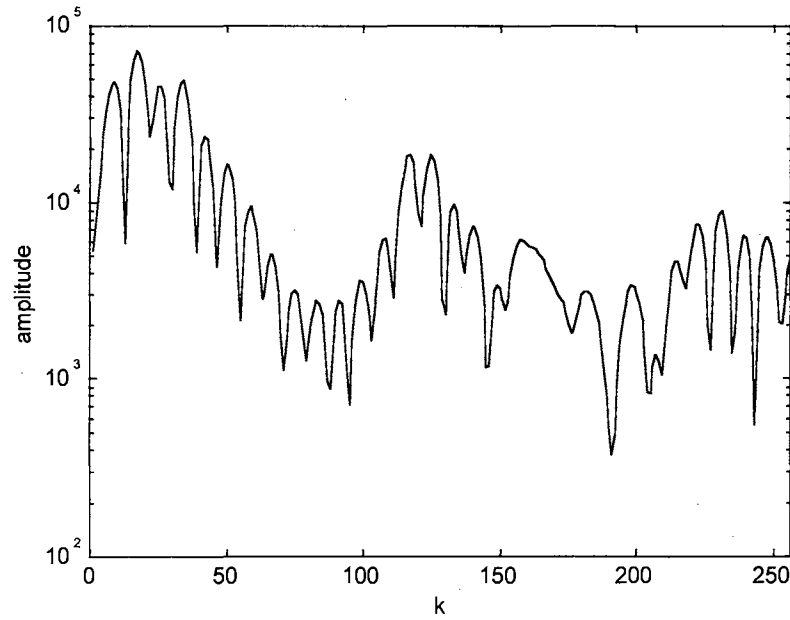


Figure 2-11. Original Spectrum for a Frame of Speech, $|S_w(\omega)|$

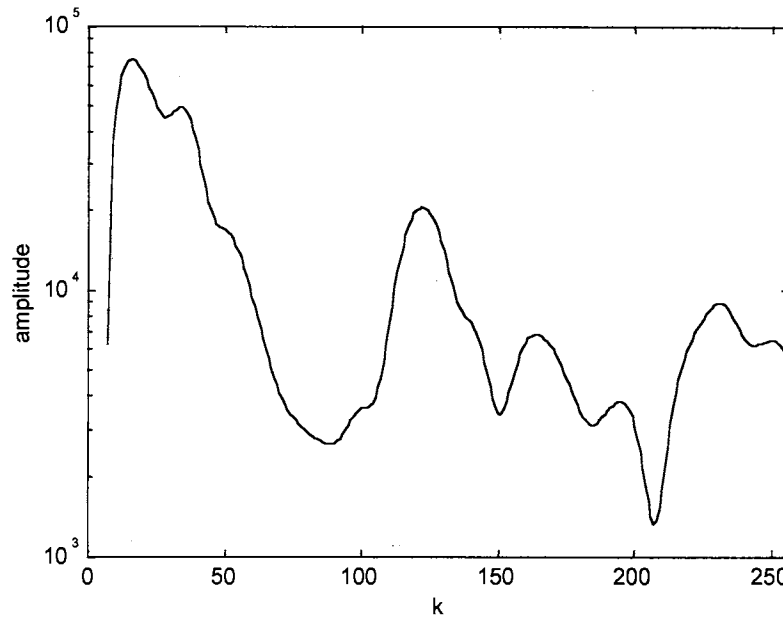


Figure 2-12. Spectral Envelope for Original Spectrum, $|H_w(\omega)|$

An all voiced (harmonic) magnitude spectrum $|P_w(\omega)|$, assumed to be flat, generated using a sinusoidal oscillator tuned to the pitch estimate and its harmonics is shown in 2-13. An unvoiced magnitude spectrum $|U_w(\omega)|$ generated using a pseudo-random white noise sequence is shown in Figure 2-14. The voiced and unvoiced decisions for the original frame of speech are provided in Figure 2-15. A value of 1 indicates that the frequency range(s) of interest is declared voiced, and a value of 0 indicates that the frequency range(s) of interest is declared unvoiced.

The mixed excitation magnitude spectrum $|E_w(\omega)|$ shown in Figure 2-16 is generated by applying the all voiced magnitude spectrum $|P_w(\omega)|$ over the ranges where the magnitude spectrum is declared voiced and by applying the unvoiced magnitude spectrum $|U_w(\omega)|$ over the range where the magnitude spectrum is declared unvoiced.

The voiced and unvoiced magnitude spectra are summed to produce the mixed excitation magnitude spectrum, $|E_w(\omega)|$. To generate the synthetic magnitude spectrum $|\tilde{S}_w(\omega)|$ shown in Figure 2-17, the mixed excitation magnitude spectrum $|E_w(\omega)|$ is multiplied by the spectral envelope magnitude spectrum $|H_w(\omega)|$ as shown in equation 2-7. This is an estimate for the reconstructed magnitude spectrum of the current frame, where the inverse Fourier transform of this spectrum is an estimate for the original input speech signal. The synthetic magnitude spectrum $|\tilde{S}_w(\omega)|$ matches the original magnitude spectrum $|S_w(\omega)|$ well in a band-based sense. This is a significant improvement on the *traditional* speech production model, which makes only a single voiced/unvoiced decision [5].

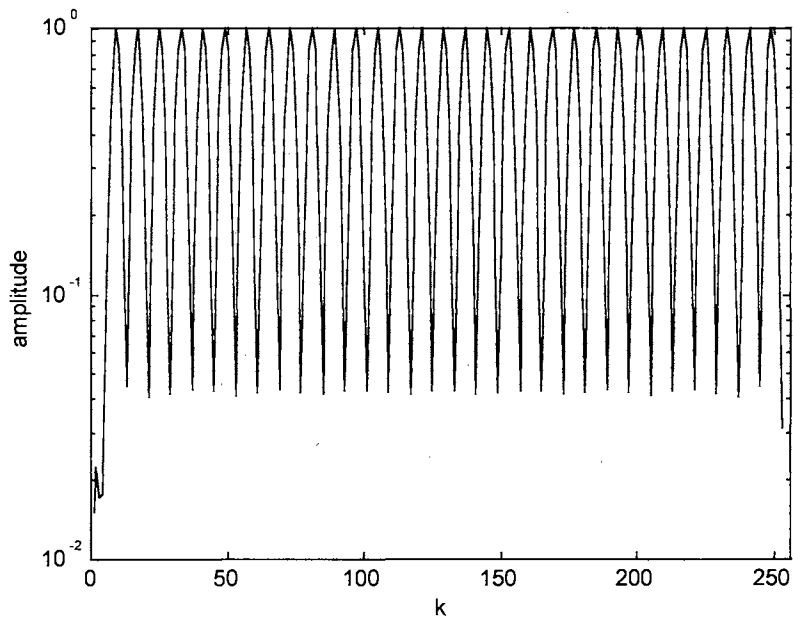


Figure 2-13. All Voiced Synthetic Spectrum, $|P_w(\omega)|$

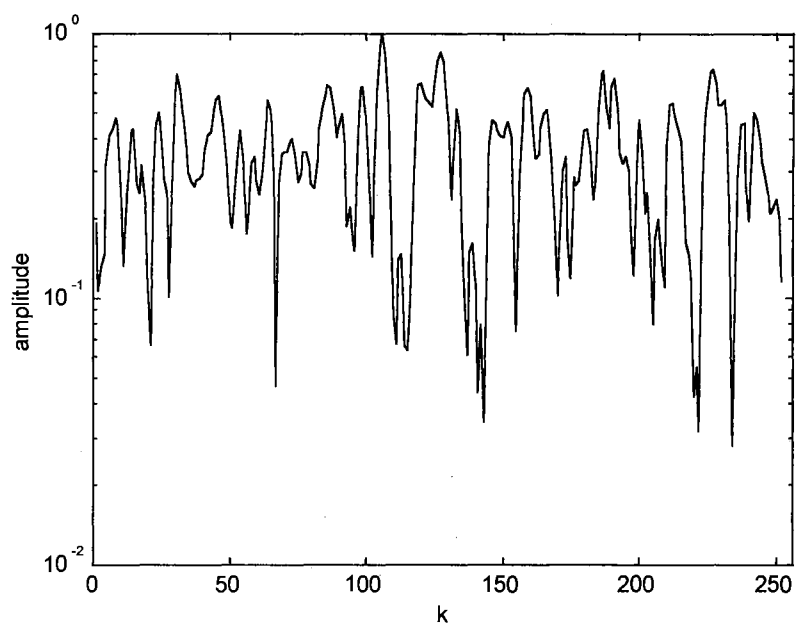


Figure 2-14. All Unvoiced Synthetic Spectrum, $|U_w(\omega)|$

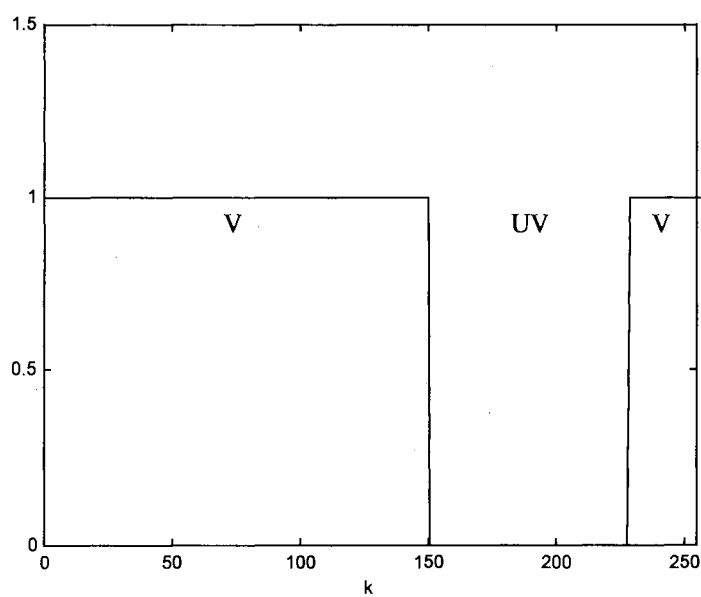


Figure 2-15. Estimated Voicing Decisions

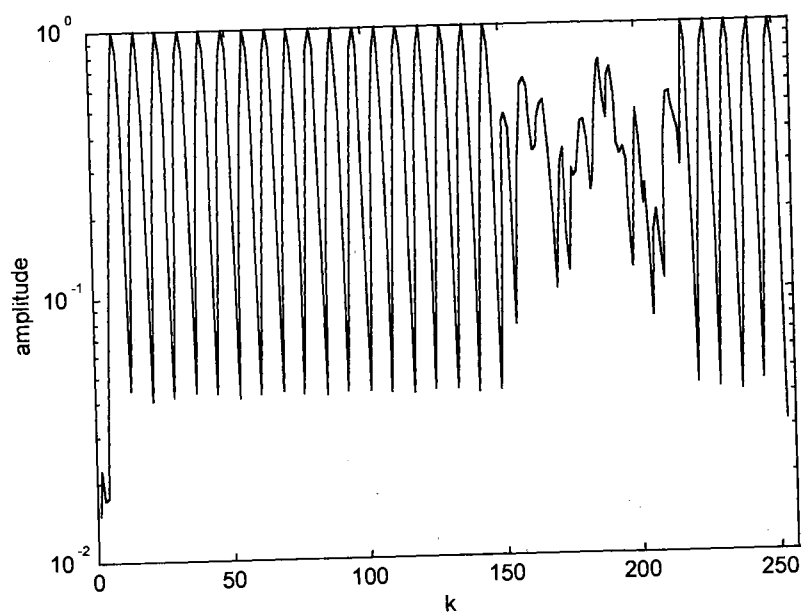


Figure 2-16. Mixed Excitation Spectrum, $|E_w(\omega)|$

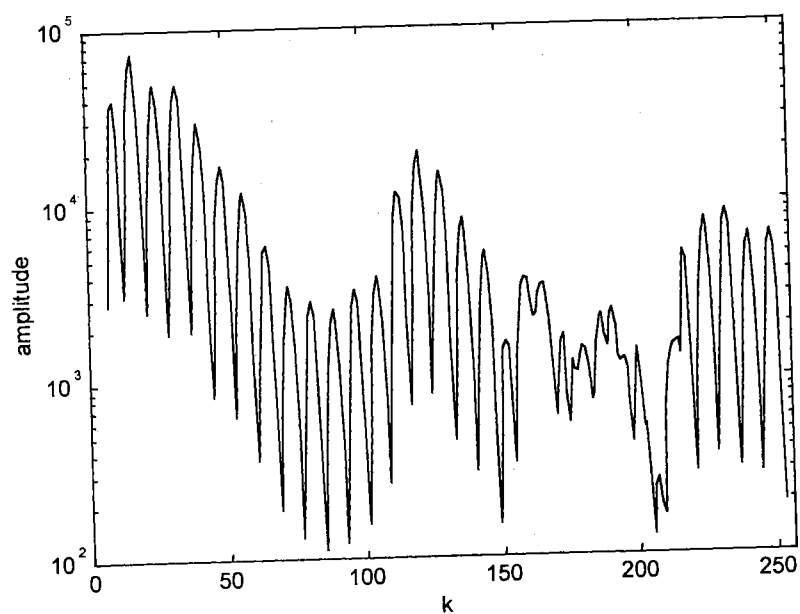


Figure 2-17. Estimated Synthetic Spectrum, $|\tilde{S}_w(\omega)|$

The practical implementation of MBE known as Improved MultiBand Excitation (IMBE) is slightly different than the concept presented. The IMBE analyzer provides a pitch estimate, multiple voiced/unvoiced decisions, and an estimate of the vocal tract response. The vocal tract response is determined by estimating the amplitudes of the pitch harmonics from the DFT magnitude spectrum for each frame of speech. Parameter estimates are made approximately every 20 ms in the analyzer. This produces a frame rate of 50 frames per second, with 96 bits per frame and a bit rate of 4,800 bps [4], [25].

The IMBE analyzer generates a parameter vector containing quantized versions of the pitch estimate, voiced/unvoiced decisions, and amplitudes of the pitch harmonics. The pitch estimate is quantized with 8 bits per frame. There are a maximum of 12 bands (actually varies with the pitch) used for the voiced/unvoiced decisions, so 12 bits are allocated for each frame. The amplitudes of the pitch harmonics are coded using the Discrete Cosine Transform (DCT) with 76 (increases as the number of bands decreases) bits being allocated for each frame [19].

In the IMBE synthesizer, combining either the appropriate voiced or unvoiced synthetic speech over each frequency band, such that the entire magnitude spectrum is covered, forms the reconstructed frame. The aggregate synthetic signal magnitude spectrum exhibits the property of mixed excitation. With MBE, the voiced bands are synthesized using sinusoidal oscillators tuned to harmonics of the estimated pitch. Each harmonic, which is declared voiced by the analyzer, is reconstructed this way. The harmonics are connected smoothly from frame to frame using equations 2-8, 2-9, and 2-10. MBE uses a linear frequency track, compared to the cubic phase interpolation of STC,

to smooth the l^{th} harmonic from the current frame to the next frame [4], [5], [26]. The current pitch $\omega_o(0)$ and the next pitch $\omega_o(S)$ are associated with $t = 0$ and $t = S$, where S represents the window shift. The initial phase ϕ_o and frequency deviation $\Delta\omega$ are chosen so that they are equal to the measured harmonic phases of the current and next frame, respectively.

$$\tilde{s}_v(t) = \sum_{l=1}^L \hat{A}_l \cos(\hat{\theta}_l(t)) \quad (2-8)$$

$$\theta_l(t) = \int_0^t \omega_l(\xi) d\xi + \phi_o \quad (2-9)$$

$$\omega_l(t) = l\omega_o(0) \frac{(S-t)}{S} + l\omega_o(S) \frac{t}{S} + \Delta\omega_l \quad (2-10)$$

Each unvoiced band is reconstructed using bandpass filtered white noise. The results of the voiced synthesis and the unvoiced synthesis are then summed producing an estimate for the current frame.

The MBE speech model has demonstrated that high quality speech is achieved at low bit rates. In fact IMBE was adopted by INMARSAT for satellite voice communications and by APCO [19], [27]. Other advantages of using MBE are robustness to additive noise and ability to be implemented in a real time system. One disadvantage to MBE is the inability to properly represent non-speech like sounds due to the assumed harmonic structure, unlike STC, which has been shown to be capable of reproducing non-speech like sounds.

The MBE speech analysis model has been shown to be capable of producing high quality speech. For this reason and others stated above, an enhanced version of MBE directed towards lower bit rates (2,400) was developed at Oklahoma State University. The enhanced version of MBE, known as EMBE, is described in the following chapter.

2.6 Summary

This chapter introduced two types of speech coders: waveform coders and voice coders. Two approaches to waveform coding; time-domain waveform coding and frequency-domain waveform coding, were discussed briefly and a number of examples were provided. The voice coder (vocoder) was discussed in more detail. The *traditional* speech production model was introduced and contrasted to several modern vocoders. These vocoders were CELP, STC, and MBE (IMBE). The CELP vocoder falls in the category of analysis-by-synthesis vocoders, and STC and MBE are often referred to as harmonic vocoders. All of the harmonic vocoders use a sinusoidal reconstruction technique for speech synthesis.

3 ENHANCED MBE

3.0 Introduction

This chapter describes the development of an Enhanced MultiBand Excitation (EMBE) speech coder at 2,400 bps. EMBE was developed at Oklahoma State University for the Department of Defense Digital Voice Processing Consortium (DDVPC) as a candidate for a new Federal Standard at 2,400 bps [6], [7]. This is the work completed over the last 4 years that led up to the developments in this dissertation.

This vocoder uses MultiBand Excitation as the speech analysis model and uses a sinusoidal model for the synthesis [4], [5], [6]. The speech is analyzed sequentially every 15 ms on 30 ms overlapping analysis frames, producing a frame rate of approximately 67 frames/second. An alternating superframe/subframe analysis strategy is applied so as to reduce the total number of parameters being produced each second, thus reducing the required bit rate. Each 30 ms superframe consists of a full update of all coder parameters, while each 15 ms subframe represents only a partial update. A full analysis and update occurs once every 30 ms. This framing approach is found to be sufficient for good temporal resolution.

Analysis consists of prefiltering, parameter estimation, quantization, and coding. Parameter decoding and frame-by-frame reconstruction of the coded speech form the synthesis stage. The relevant parameters that are used to represent the input speech

waveform, are fundamental frequency (pitch), vocal tract spectrum, voicing decisions, and gain. A basic block diagram of the analyzer is shown in Figure 3-1, and the synthesizer is shown in Figure 3-2.

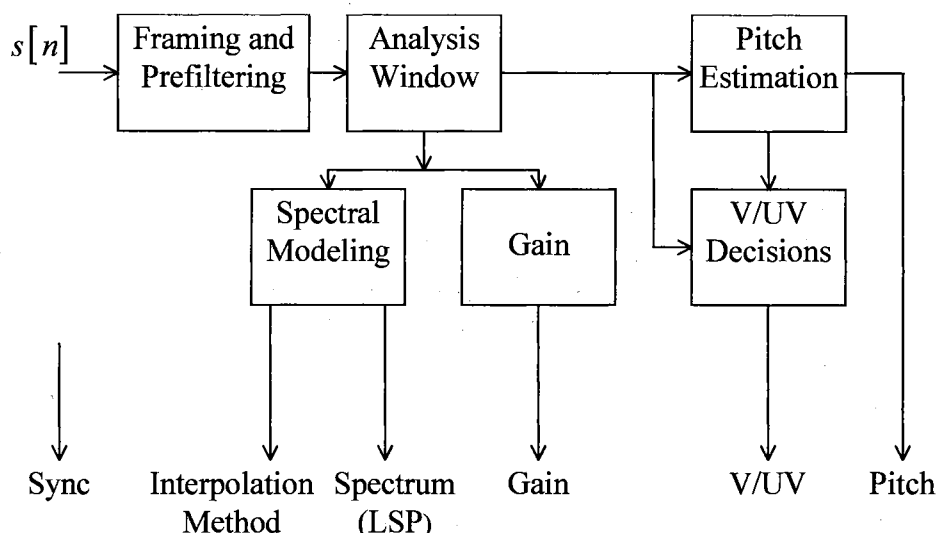


Figure 3-1. Block Diagram of EMBE Analyzer

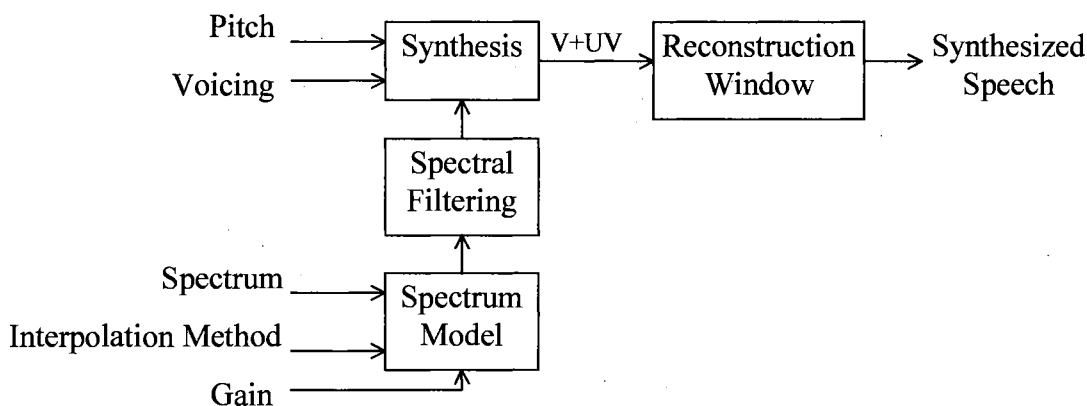


Figure 3-2. Block Diagram of EMBE Synthesizer

The following sections describe the procedures used to estimate, quantize, and code the relevant parameters into a 2,400 bps bit stream, decode the coded parameters from the bit stream, and reconstruct high quality speech from the decoded parameters. It

is assumed that the reader is familiar with short-time analysis, so the details of the implementation are not presented.

3.1 Analyzer

3.1.0 Introduction

The EMBE analyzer estimates the following parameters: pitch, voicing, spectrum, and gain. These parameters are quantized and coded for either transmission or storage. A basic block diagram of the EMBE analyzer is given in Figure 3-1. The input speech is framed, filtered and windowed into two separate data paths. The pitch estimate and the voicing decisions are computed in one path while the spectrum and gain are computed in the other path. The following sections describe the implementation used in the development of a 2,400 bps Enhanced MultiBand Excitation Vocoder.

3.1.1 Pre-filtering and Windowing

The input speech, $s[n]$, is filtered with a high pass filter with a cutoff frequency of approximately 70 Hz. The frequency response for this filter is provided in Figure 3-3.

This filter is used mainly for removing the low frequency components that may inhibit the parameter estimation. For example, the pitch is only estimated over the range 70 Hz to 400 Hz, so frequencies below 70 Hz are not needed for analysis.

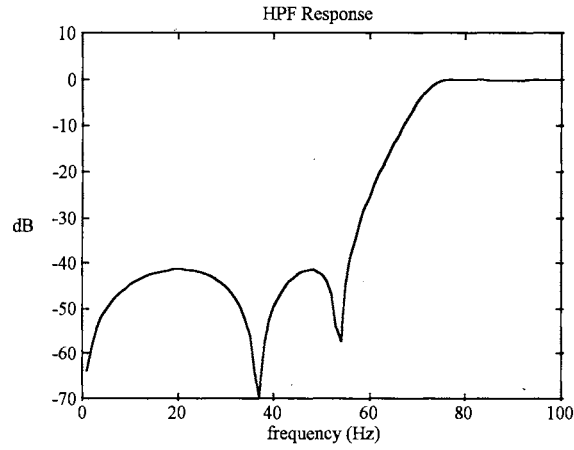


Figure 3-3. Frequency Response for Input HPF

The high pass filter is a 5th order elliptic filter with 0.25 dB of ripple in the passband and more than 20 dB of attenuation at 60 Hz. The filter transfer function, with quantized coefficients, is shown below in equations 3-1 and 3-2.

$$H_{HP}(z) = \frac{(b_{11} + b_{12}z^{-1})(b_{21} + b_{22}z^{-1} + b_{23}z^{-2})(b_{31} + b_{32}z^{-1} + b_{33}z^{-2})}{(a_{11} + a_{12}z^{-1})(a_{21} + a_{22}z^{-1} + a_{23}z^{-2})(a_{31} + a_{32}z^{-1} + a_{33}z^{-2})} \quad (3-1)$$

$$b_{11} = b_{21} = b_{23} = b_{31} = b_{33} = 0.9712596$$

$$b_{12} = -b_{11}$$

$$b_{22} = -1.941714$$

$$b_{32} = -1.940796$$

$$a_{11} = a_{21} = a_{31} = 1.0 \quad (3-2)$$

$$a_{12} = -0.8971591$$

$$a_{22} = -1.987283$$

$$a_{23} = 0.9905484$$

$$a_{32} = -1.938817$$

$$a_{33} = 0.9437869$$

The high pass filtered signal, $s_{HP}[n]$, is computed using

$$s_{HP}[n] = \sum_{r=0}^{N-1} s[r] h_{HP}[n-r] \quad (3-3)$$

where N is the amount of data to be filtered, $s[n]$ is the input speech signal, $h_{HP}[n]$ is the impulse response for the transfer function $H_{HP}(z)$, and n ranges from 0 to $N-1$.

After high pass filtering the input speech, it is windowed producing the input frames. A rectangular window is used for determining a coarse pitch estimate, a square-root of Hamming window is used for pitch refinement, and a Hamming window is used to perform spectral analysis. The windowing operations are computed from

$$s_{p_c}[n] = s_{HP}[n] w_R[n] \quad (3-4)$$

$$s_{p_f}[n] = s_{HP}[n] w_{SQORTH}[n] \quad (3-5)$$

$$s_s[n] = s_{HP}[n] w_H[n], \quad (3-6)$$

where $s_{p_c}[n]$ represents the data used to estimate the coarse pitch estimate, $s_{p_f}[n]$ represents the data used in refining the coarse pitch estimate, and $s_s[n]$ represents the data used for computing the spectrum. The three windows $w_R[n]$, $w_{SQORTH}[n]$, and $w_H[n]$ are defined below in equations 3-7, 3-8, and 3-9.

$$w_R[n] = \begin{cases} 1 & 0 \leq n \leq N_a - 1 \\ 0 & \text{otw} \end{cases} \quad (3-7)$$

$$w_{SQORTH}[n] = \begin{cases} \left(0.54 - 0.46 \cos\left[\frac{2\pi n}{N_a - 1}\right] \right)^{\frac{1}{2}} & 0 \leq n \leq N_a - 1 \\ 0 & \text{otw} \end{cases} \quad (3-8)$$

$$w_H[n] = \begin{cases} 0.54 - 0.46 \cos\left[\frac{2\pi n}{N_a - 1}\right] & 0 \leq n \leq N_a - 1 \\ 0 & \text{otw} \end{cases} \quad (3-9)$$

where the variable N_a defines the length of the analysis windows.

For convenience, the magnitude spectra, $S_s(k)$ and $S(k)$, of $s_s[n]$ and $s_{p_f}[n]$ are computed using a M length DFT as defined in equations 3-10 and 3-11. These definitions are used throughout this chapter.

$$S_s(k) = \left| \sum_{n=0}^{M-1} s_s[n] e^{-j\frac{2\pi nk}{M}} \right| \quad (3-10)$$

$$S_p(k) = \left| \sum_{n=0}^{M-1} s_{p_f}[n] e^{-j\frac{2\pi nk}{M}} \right| \quad (3-11)$$

3.1.2 Pitch Estimate

The analyzer forms a refined pitch estimate twice for each analysis frame (one for each subframe). MBE-based vocoders require a high degree of pitch accuracy for analysis, so a two-stage estimation procedure is used. In the first stage, a coarse pitch estimate P_c is computed using a procedure based on the simple inverse filter tracking (SIFT) pitch detector [30]. The block diagram for the coarse pitch estimate is shown in Figure 3-4. The pitch is assumed to lie in the range from 20 samples (400 Hz) to 114 samples (70 Hz). A pitch falling outside this interval is considered to be incorrect and no pitch is computed, thus indicating an unvoiced frame.

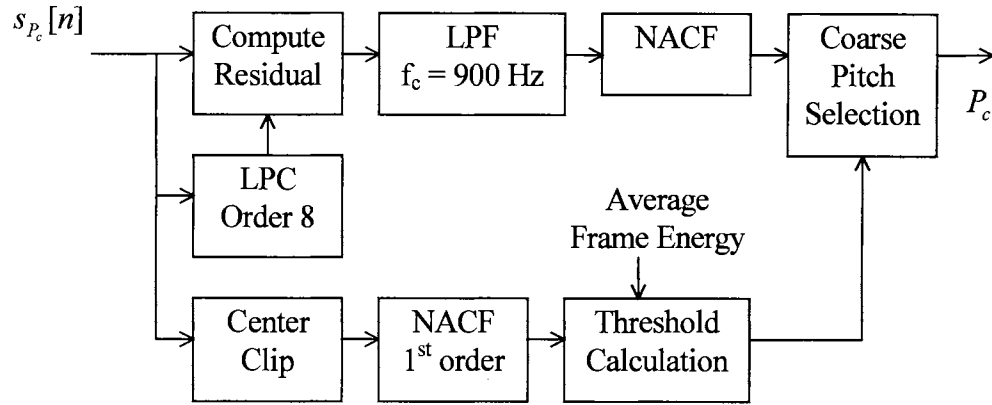


Figure 3-4. Block Diagram for Initial Pitch Estimate

The residual $r[m]$ of an all-pole model is computed from

$$r[m] = s_{P_c}[m] - \sum_{n=1}^p \alpha_n s_{P_c}[m-n], \quad (3-12)$$

where $p = 8$ is the order of the linear predictor. The linear predictor coefficients, α_n 's, are solved using the standard Levinson-Durbin Recursion algorithm with a pre-whitening factor of $\frac{1}{24576}$ applied to the zeroth autocorrelation coefficient. The pre-whitening factor ensures that the autocorrelation matrix does not become ill conditioned. The residual is then bandlimited to approximately 900 Hz using a 6th order elliptic low pass filter with 0.25 dB of ripple in the passband and more than 40 dB of attenuation at 1,000 Hz. The frequency response for the low pass filter is shown below in Figure 3-5 and described in equations 3-13 and 3-14.

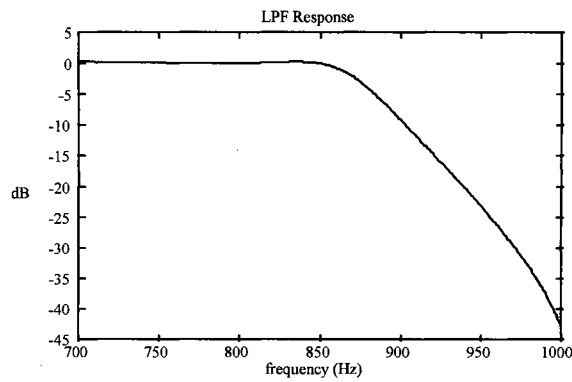


Figure 3-5. Frequency Response of Residual LPF

$$H_{LP}(z) = \frac{(b_{11} + b_{12}z^{-1} + b_{13}z^{-2})(b_{21} + b_{22}z^{-1} + b_{23}z^{-2})(b_{31} + b_{32}z^{-1} + b_{33}z^{-2})}{(a_{11} + a_{12}z^{-1} + a_{13}z^{-2})(a_{21} + a_{22}z^{-1} + a_{23}z^{-2})(a_{31} + a_{32}z^{-1} + a_{33}z^{-2})} \quad (3-13)$$

$$b_{11} = b_{13} = b_{21} = b_{23} = b_{31} = b_{33} = 0.2404547$$

$$b_{12} = -0.3355389$$

$$b_{22} = -0.2791540$$

$$b_{32} = -0.1008389$$

$$a_{11} = a_{21} = a_{31} = 1.0 \quad (3-14)$$

$$a_{12} = -1.419309$$

$$a_{13} = 0.5453778$$

$$a_{22} = -1.464743$$

$$a_{23} = 0.7784000$$

$$a_{32} = -1.512890$$

$$a_{33} = 0.9443823$$

The low pass filtered signal, $r_{LP}[n]$, is computed from

$$r_{LP}[n] = \sum_{i=0}^{N_a-1} r[i] h_{LP}[n-i], \quad (3-15)$$

where $r[n]$ is the residual signal and $h_{LP}[n]$ is the impulse response of the transfer function $H_{LP}(z)$.

After the residual is bandlimited, the normalized autocorrelation function, $R(j)$, is computed using equation 3-16. The coarse pitch estimate P_c is then chosen as the index corresponding to the maximum value in $R(j)$ based on a set of heuristic criteria as described in the following paragraphs.

$$R(j) = \frac{\sum_{k=1}^{N-1} r_{LP}[k] r_{LP}[j+k]}{\sum_{k=0}^{N-1} r_{LP}[k]} \quad (3-16)$$

A correlation threshold τ_p is developed to determine if a peak exists in $R(j)$ that is assumed to correspond to the pitch for a given frame. This threshold is developed based on the energy in the current frame E_f and energy threshold E_{uv} (referred to as unvoiced energy threshold), which are computed as shown

$$E_f = \sum_{n=0}^{N-1} s_{p_c}^2[n] \quad (3-17)$$

$$E_{uv} = \begin{cases} 0.995E_{uv} + 0.005E_f & 0.75E_{uv} \geq 10^4 \\ \frac{10^4}{0.75} & 0.75E_{uv} < 10^4 \end{cases} \quad (3-18)$$

where 10^4 represents a lower bound on the frame energy. A second criterion used for developing the correlation threshold is based on the first order normalized autocorrelation

coefficient. The amplitude of this coefficient assists in determining frames that are clearly unvoiced. This coefficient is computed using the following method.

First, the input data $s_p[n]$ is center-clipped as shown in equations 3-19 through 3-22.

$$s_c[n] = \begin{cases} s_p[n] - C_L & s_p[n] > C_L \\ s_p[n] + C_L & s_p[n] < -C_L \\ 0 & \text{otw} \end{cases} \quad (3-19)$$

$$C_L = 0.60 \min\{a, b\} \quad (3-20)$$

$$a = \max\{s_p[n]\} \quad 0 < n < \left(\frac{N_a}{3} - 1\right) \quad (3-21)$$

$$b = \max\{s_p[n]\} \quad \left(N_a - \frac{N_a}{3}\right) < n < N_a - 1 \quad (3-22)$$

The first order normalized autocorrelation $R'(1)$ is then determined using the center-clipped data $s_c[n]$ as given by

$$R'(1) = \frac{\sum_{k=1}^{N-1} s_c[k] s_c[1+k]}{\sum_{k=0}^{N-1} s_c[k]}. \quad (3-23)$$

Now based on the first order normalized autocorrelation $R'(1)$ and the frame energy E_f the correlation threshold τ_p is defined by

$$\tau_p = \begin{cases} 0.55 & R'(1) < 0.25 \\ \lambda & \text{otw} \end{cases} \quad (3-24)$$

$$\lambda = \begin{cases} 0.325 & E_f \geq 0.75 E_{uv} \\ 0.65 & E_f < 0.75 E_{uv} \end{cases}. \quad (3-25)$$

The coarse pitch estimate P_c is then determined based on the correlation threshold τ_p as given by

$$P_c = \begin{cases} j_{\max(R(j))} & j_{\max(R(j))} > \tau_p \quad E_f > 10^5 \\ 0 & \text{otw} \end{cases} \quad (3-26)$$

The coarse pitch estimate P_c is then tested to determine if it is actually a subharmonic, multiple, or the real pitch. This in essence is a check for either doubling or halving of the current frame's pitch estimate. This is accomplished by searching the autocorrelation $R(j)$ for other peaks, \hat{P}_{c_i} , that meet the amplitude requirement as given in equation 3-27 and are determined to be a submultiple. The coarse pitch estimate P_c for the current analysis frame is chosen as the minimum of all the possible candidates as shown in equation 3-28. This coarse pitch estimate P_c is only accurate to within one sample. MBE based speech coders require pitch estimates with fractional sample accuracy. This accuracy is achieved by adding a pitch refinement stage. The refinement stage is discussed in the following paragraphs.

$$\hat{P}_{c_i} = \frac{j}{i} \quad R\left(\frac{j}{i}\right) \geq 0.55R(P_c) \quad i = 2,3,4 \quad (3-27)$$

$$P_c = \min\{P_c, \hat{P}_{c_i}\} \quad (3-28)$$

3.1.3 Pitch Refinement

Since MBE-based vocoders are dependent on accurate pitches, a second stage is added to refine the coarse estimate to sub-sample accuracy. A block diagram of the pitch

refinement is shown in Figure 3-6 below. This stage uses a frequency domain interpolation and error minimization to perform the refinement, based on analysis-by-synthesis. An all voiced synthetic spectrum is built based on the pitch estimate from the first stage. This synthetic spectrum is then matched to the original spectrum. Determination of the final pitch estimate P_f is performed in two steps. The first step uses look ahead and look back to determine a coarse pitch estimate to refine. A check for doubling and halving is performed, and a smoothness constraint is imposed. The second step is the actual refinement of the coarse pitch estimate.

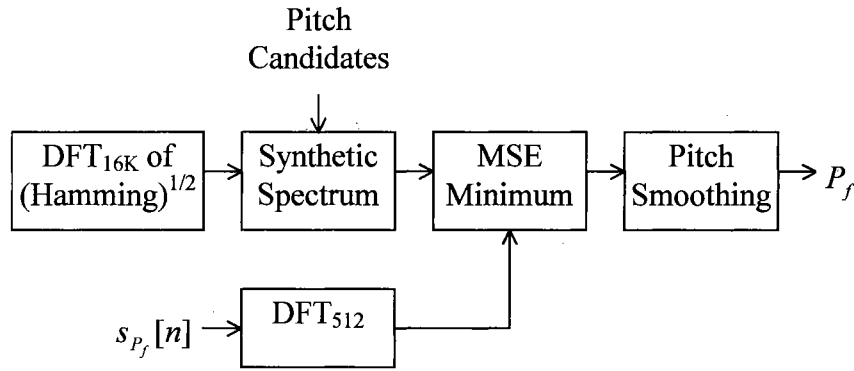


Figure 3-6. Block Diagram for Pitch Refinement

In the first step four candidate pitch estimates are input to the analysis-by-synthesis routine. These candidate pitches are determined as shown below in equations 3-29 through 3-31. P_{f-1} represents the refined pitch estimate in the previous frame, P_{c_0} represents the coarse pitch estimate for the current frame, $P_{c_{-1}}$ corresponds to the coarse pitch estimate in the previous frame, and $P_{c_{+1}}$ corresponds to the coarse pitch estimate in the future frame.

$$\{\tilde{P}_1, \tilde{P}_2, \tilde{P}_3, \tilde{P}_4\} \quad (3-29)$$

$$\tilde{P}_1 = P_{f_{-1}} \quad \tilde{P}_2 = P_{c_0} \quad (3-30)$$

$$\tilde{P}_3 = \begin{cases} P_{c_{-1}} & R(P_{c_0}) \geq 0.625 \\ 0.5P_{c_0} & \text{otw} \end{cases} \quad \tilde{P}_4 = \begin{cases} P_{c_{+1}} & R(P_{c_0}) \geq 0.625 \\ 2P_{c_0} & \text{otw} \end{cases} \quad (3-31)$$

Next, an all voiced synthetic magnitude spectrum, $\tilde{S}_{\omega_i}(k)$, is generated for each candidate pitch listed above in equation 3-29. The four pitch candidates are determined from equations 3-30 and 3-31. The all voiced magnitude spectrum is computed using equation 3-32, where A_l is the amplitude of the harmonic being generated and H represents the magnitude spectrum for a square-root of Hamming window. Equations 3-33 and 3-34 are used to determine A_l and H , respectively. The upper and lower limits for a particular harmonic are found using equation 3-35. The candidate pitches ω_i , in radians, are computed using equation 3-36 and the number of harmonics L_i corresponding to each pitch candidate are found using equation 3-37.

$$\tilde{S}_{\omega_i}(k) = A_l H \left(\left\lfloor \frac{16384k}{M} - \frac{16384\omega_i l}{2\pi} + 0.5 \right\rfloor \right) \quad 1 \leq l \leq L_i \quad (3-32)$$

$$A_l = \frac{\sum_{k=lower}^{upper-1} S_p(k) H \left(\left\lfloor \frac{16384k}{M} - \frac{16384\omega_i l}{2\pi} + 0.5 \right\rfloor \right)}{\sum_{k=lower}^{upper-1} \left[H \left(\left\lfloor \frac{16384k}{M} - \frac{16384\omega_i l}{2\pi} + 0.5 \right\rfloor \right) \right]^2} \quad (3-33)$$

$$H(k) = \left| \sum_{n=0}^{16383} w_H[n] e^{-j \frac{2\pi n k}{16384}} \right| \quad (3-34)$$

$$upper = \left\lceil \frac{M}{2\pi} (l + 0.5) \omega_i \right\rceil \quad lower = \left\lceil \frac{M}{2\pi} (l - 0.5) \omega_i \right\rceil \quad (3-35)$$

$$\omega_i = \frac{2\pi}{\tilde{P}_i} \quad i = 1, 2, 3, 4 \quad (3-36)$$

$$L_i = \left\lfloor \frac{0.975\pi}{\omega_i} \right\rfloor \quad (3-37)$$

After the all voiced synthetic magnitude spectrum is generated, an error value ε_i is computed over the first 40% of the frequency spectrum as given by

$$\varepsilon_i = \frac{1}{\kappa} \sum_{k=1}^{0.4L_i} \left(S_p(k) - \tilde{S}_{\omega_i}(k) \right)^2 \quad i = 1, 2, 3, 4. \quad (3-38)$$

The error ε_i varies in a non-linear fashion with frequency, so a bias term κ is needed to normalize the effect. The bias is computed as a function of the ending frequency value, $0.4L_i$ in this case, and a polynomial that approximates the non-linearly varying error curve [4], [6]. The computation of the bias coefficient κ is computed using equations 3-39 and 3-40.

$$\kappa = 0.4L_i \left(a_4 x^4 + a_3 x^3 + a_2 x^2 + a_1 x + a_0 \right) \quad (3-39)$$

$$a_0 = -7.40897(10^{-1})$$

$$a_1 = 1.88602(10^{-2})$$

$$a_2 = -9.71097(10^{-5}) \quad (3-40)$$

$$a_3 = 2.31857(10^{-7})$$

$$a_4 = -2.07128(10^{-10})$$

$$x = \tilde{P}_i$$

The coarse pitch estimate, \hat{P} sent to the pitch refinement stage, is determined by the criteria of

$$\hat{P} = \begin{cases} \tilde{P}_{\min(\epsilon_i)} & \min(\epsilon_i) \leq 0.825 \\ 0 & \text{otw} \end{cases} \quad (3-41)$$

In the second step, given a valid pitch, the pitch estimate \hat{P} is varied plus and minus in quarter-sample intervals resulting in a new set of pitch candidates \tilde{P}_j shown below in equations 3-42 and 3-43.

$$\{\tilde{P}_1 \quad \tilde{P}_2 \quad \tilde{P}_3 \quad \tilde{P}_4 \quad \tilde{P}_5 \quad \tilde{P}_6 \quad \tilde{P}_7\} \quad (3-42)$$

$$\begin{aligned} \tilde{P}_1 &= \hat{P} - 0.75 \\ \tilde{P}_2 &= \hat{P} - 0.50 \\ \tilde{P}_3 &= \hat{P} - 0.25 \\ \tilde{P}_4 &= \hat{P} \\ \tilde{P}_5 &= \hat{P} + 0.25 \\ \tilde{P}_6 &= \hat{P} + 0.50 \\ \tilde{P}_7 &= \hat{P} + 0.75 \end{aligned} \quad (3-43)$$

Again, an all voiced synthetic magnitude spectrum is generated per equations 3-32 through 3-37 for each \tilde{P}_j . The spectral error ϵ_i is now computed over 80% of the frequency spectrum, instead of the 40% in the first stage, as given by

$$\epsilon_i = \frac{1}{K} \sum_{k=1}^{0.8L_j} \left(S_p(k) - \tilde{S}_i(k) \right)^2 \quad i = 1, 2, 3, 4, 5, 6, 7. \quad (3-44)$$

The subsample pitch estimate corresponding to the synthetic spectrum producing the lowest spectral error ϵ_i is chosen as the refined pitch, P_f , for the current frame as shown in equation 3-45. This pitch is used for estimating the remaining parameters, such as voicing decisions, spectral envelope, and gain.

$$P_f = \tilde{P}_{\min(\epsilon_i)} \quad (3-45)$$

3.1.4 Voicing

The voiced and unvoiced decisions are the heart of the MBE analysis model. It is assumed that the speech spectrum is composed of both voiced and unvoiced bands. This is equivalent to considering the excitation to contain both periodic and aperiodic components simultaneously. A block diagram for estimating the voicing decisions is provided in Figure 3-7.

A voicing decision is formed for each harmonic of the fundamental using a frequency domain procedure similar to the pitch refinement stage. The synthetic spectrum corresponding to the refined pitch estimate is matched to the original spectrum. An error term is computed for each harmonic of the synthetic spectrum. This error term determines whether the match for a given harmonic is 'good' (low error) or not. This error value is compared to an adaptive threshold function that is determined by the pitch, the voicing decision, and the harmonic. The voicing decisions are then grouped into four non-linear bands covering the entire speech spectrum. A single voicing decision is made for each band based on the individual harmonic-based voicing decisions. The band structure is variable based on pitch and human perception - the most resolution and accuracy are maintained in the lower frequency bands.

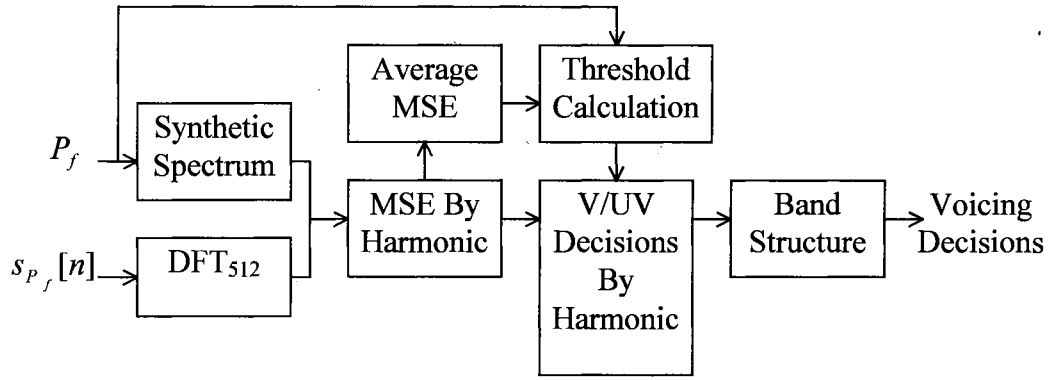


Figure 3-7. Block Diagram for Estimating Voicing Decisions

The voicing decisions are estimated by computing a spectral harmonic error, ε_l , for each harmonic corresponding to the all voiced synthetic magnitude spectrum generated using the final pitch estimate P_f for the current frame. The error term is calculated via

$$\varepsilon_l = \frac{1}{\kappa} \sum_{k=lower}^{upper-1} \frac{(S_s(k) - \tilde{S}_{P_f}(k))^2}{S_s(k)^2} \quad 1 \leq l \leq L, \quad (3-46)$$

where κ is defined in equation 3-39, except $x = P_f$, and L is defined by equation 3-37.

The average error over the lower half of the frequency spectrum is computed using equation 3-47 below. The voicing threshold function T_l is computed using equations 3-48 and 3-49. This threshold function is computed for every potential harmonic in the frequency spectrum.

$$\varepsilon_{\frac{L}{2}} = \frac{2}{L} \sum_{l=1}^{\frac{L}{2}} \varepsilon_l \quad (3-47)$$

$$T_l = \begin{cases} \frac{0.8}{1 + 0.3\varepsilon_{\frac{L}{2}}} & l < \eta \\ \frac{0.8}{1 + 0.3\varepsilon_{\frac{L}{2}}} \min\left(1, \frac{0.9(\eta-1) + 0.1l - L}{(\eta-1-L)}\right) & l \geq \eta \end{cases} \quad (3-48)$$

$$\eta = \min(9, l) \quad \varepsilon_l > \varepsilon_{\frac{L}{2}} \quad (3-49)$$

The harmonic spectral error ε_l is compared to the voicing threshold and each harmonic is declared either voiced or unvoiced based on the criteria given by

$$V_l = \begin{cases} 1 & \varepsilon_l < T_l \\ 0 & \text{otw} \end{cases}, \quad (3-50)$$

where a '1' corresponds to a voiced harmonic and '0' corresponds to an unvoiced harmonic. After the individual harmonic voicing decisions V_l are made, the frequency spectrum is split into subbands. A non-linear band structure B_L , using 4 bands, is defined by

$$B_L = \begin{cases} 5 & 11 & 15 & 16-25 & L \geq 42 \\ 5 & 9 & 13 & 5-14 & 32 \leq L < 42 \\ 5 & 7 & 7 & 6-12 & 25 \leq L < 32, \\ 3 & 5 & 5 & 2-11 & 15 \leq L < 25 \\ 3 & 3 & 3 & 0-6 & L < 15 \end{cases}, \quad (3-51)$$

where L represents the number of harmonics in the current frame.

The voicing decisions for each band are determined using a majority function. If a majority of the harmonics in a given band has been declared voiced then the band is declared voiced, otherwise the band is declared unvoiced. The first three bands are determined using the majority rule. The last band requires a slightly different approach,

since the number of harmonic varies from frame to frame. Again, if a majority of the harmonics has been declared voiced the last band is declared voiced, otherwise the last band is declared unvoiced. The special case occurs when only one harmonic exists in the last band. In this case, the voicing decision for the last band is determined by band three. When band three is declared voiced, the harmonic in the last band is declared voiced. If band three is declared unvoiced, the last band is declared unvoiced.

As with pitch, some heuristic criteria have been applied to the voicing decisions for smoothing the decisions. If neither one of the first two bands have been declared voiced then the entire frame is declared unvoiced. The refined pitch estimate, P_f , and the voicing decisions V_l are used in estimating the two remaining parameters spectrum and gain.

3.1.5 Spectrum

The goal of a spectral model for a harmonic coder, such as EMBE, is to represent accurately the harmonic amplitudes for voiced speech, and to fit the spectrum in an average sense for unvoiced speech. Harmonic coders usually employ some direct form of quantization of the harmonic amplitudes to achieve this. While this results in a highly accurate representation of the spectrum, the number of bits required precludes its use for low bit rate coding. This is overcome with the use of a parametric model for the spectrum. The EMBE speech coder represents the spectrum using a spline enhanced, linear predictive (LP) model for voiced speech, and a traditional LP model for unvoiced speech. Both methods use a spectral warping function prior to the actual LP model computation.

Figure 3-8 below illustrates the entire spectral modeling procedure. The LP model coefficients are computed using a frequency domain approach, rather than the traditional time domain approach. This allows the manipulation of the spectrum prior to model computation to enhance perceptually important areas.

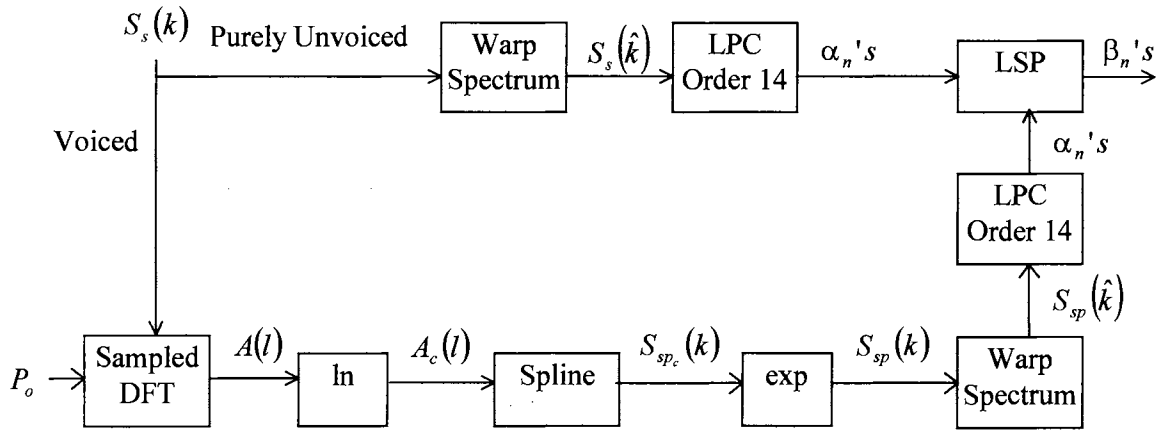


Figure 3-8. Block Diagram of Spectral Modeling

For the case of modeling unvoiced speech, the DFT of the frame, $S_s(k)$, is warped to the Mel scale to enhance the perceptually more important, lower frequency regions. This warping is given by equations 3-52 through 3-55.

$$\hat{k} = \begin{cases} k & 0 \leq k < \frac{M}{8} \\ 1000 \log_2 \left(1 + \frac{8k}{M} \right) \frac{M}{F_s} & \frac{M}{8} + 1 \leq k < \frac{M}{2} - 1 \end{cases} \quad (3-52)$$

$$M' = 2 \left\lceil 1000 \log_2 \left(1 + \frac{8 \left(\frac{M}{2} - 1 \right)}{M} \right) \frac{M}{F_s} \right\rceil \quad (3-53)$$

$$S_s(\hat{k}) = S_s(k) \Big|_{k=\hat{k}} \quad 0 \leq \hat{k} < \frac{M}{2} - 1 \quad (3-54)$$

$$S_s(\hat{k}) = S_s(M' - \hat{k}) \quad \frac{M'}{2} \leq \hat{k} < M' \quad (3-55)$$

The Mel warped DFT index is represented by \hat{k} , M' is the length of the warped DFT, and F_s refers to the sampling frequency.

For unvoiced speech, the spectrum is frequency warped to the Mel scale and then a linear predictive model is fitted to the warped data. The general LP model is given in equation 3-56 where, α_k represents the k^{th} predictor coefficient, p is the model order, and G corresponds to the model gain. A 14th order model, $p = 14$, is used to represent the spectrum. The solution for the α_k 's is computed in the frequency domain, based upon the Mel warped spectrum given in equations 3-54 and 3-55. The solution of the predictor coefficients is given by equations 3-57 through 3-59. The calculation of the LP model in the frequency domain assumes that the power spectrum is an even function of frequency and correctly sampled around the unit circle. This assumption is verified by equation 3-55. The solution to equation 3-59 is obtained using the Levinson-Durbin recursion algorithm [31], [32]. Prior to this solution, a small pre-whitening factor, mentioned earlier, is added to the zeroth correlation coefficient to ensure that the LP model solution does not become ill conditioned, as shown in equation 3-58.

$$S_{lp}(\hat{k}) = \frac{G}{1 + \sum_{n=1}^p \alpha_n e^{-\frac{2\pi \hat{k} n}{M'}}} \quad 0 \leq \hat{k} < M' \quad (3-56)$$

$$R_i = \frac{1}{M'} \sum_{k=0}^{\frac{M'}{2}-1} S_s^2(\hat{k}) \cos\left(\frac{2\pi \hat{k} i}{M'}\right) \quad 0 \leq i \leq p \quad (3-57)$$

$$R_0 = R_0 \left(1 + \frac{1}{24576} \right) \quad (3-58)$$

$$\sum_{k=1}^p \alpha_k R_{|n-k|} = -R_n \quad 1 \leq n \leq p \quad (3-59)$$

Referring again to Figure 3-8 for the case of voiced speech, the DFT for the frame is sampled at the harmonics to obtain a spectrum composed of harmonic amplitudes only. This sampling is shown in equation 3-60. A compression function is then applied to the harmonic amplitudes to reduce their dynamic range. It has been reported in [26] that the use of a compression function prior to spectral interpolation, in this case by the cubic spline, improves the spectral fit obtained through linear prediction. The form of this compression function is given in equation 3-61 below.

$$A(l) = S_s(k) \Big|_{k=\left\lfloor \frac{lM}{P_f} + 0.5 \right\rfloor} \quad 1 \leq l \leq L \quad (3-60)$$

$$A_c(l) = \ln[A(l) + 1] \quad 1 \leq l \leq L \quad (3-61)$$

Once the harmonic amplitudes have been compressed, a cubic spline envelope is fitted to the harmonic amplitudes. The cubic spline function serves to smoothly interpolate between the harmonic amplitudes to produce a slower varying curve that linear prediction is able to more accurately model. The spline equations are given in equations 3-63 through 3-66 below. These equations represent the constraints on the general cubic polynomial given in equation 3-62. These constraints are chosen to enforce continuity and smoothness at the polynomial boundaries. Equation 3-67 results from expressing the general spline equation in 3-62 in an alternate form and enforcing smoothness in its first derivative [33]. The unknowns in the equations include the spline coefficients a_i , b_i , c_i , and d_i , as well as the second derivatives of the each spline segment,

x_i . Two more conditions are needed to solve this system of equations, namely the conditions on the 1st and L^{th} spline segments. These are shown below in equations 3-68 and 3-69. The actual solution to this system of equations is given in [33].

$$S_{spc_i}(l) = a_i l^3 + b_i l^2 + c_i l + d_i \quad 1 \leq i \leq L \quad (3-62)$$

$$S_{spc_i}(l_i) = a_i l_i^3 + b_i l_i^2 + c_i l_i + d_i = A_c(l_i) \quad (3-63)$$

$$S_{spc_i}(l_{i+1}) = a_i l_{i+1}^3 + b_i l_{i+1}^2 + c_i l_{i+1} + d_{i+1} = A_c(l_{i+1}) \quad (3-64)$$

$$S_{spc_i}''(l_i) = 6a_i l_i + 2b_i l_i^2 = x_i \quad (3-65)$$

$$S_{spc_i}''(l_{i+1}) = 6a_i l_{i+1} + 2b_i l_{i+1}^2 = x_{i+1} \quad (3-66)$$

$$(l_i - l_{i-1})d_{i-1} + 2(l_{i+1} - l_{i-1})d_i + (l_{i+1} - l_i)d_{i+1} = \quad (3-67)$$

$$6 \left[\frac{A_c(l_{i+1}) - A_c(l_i)}{l_{i+1} - l_i} - \frac{A_c(l_i) - A_c(l_{i-1})}{l_i - l_{i-1}} \right]$$

$$x_1 = 0 \quad (3-68)$$

$$x_L = 0 \quad (3-69)$$

Following the computation of the spline envelope, $S_{spc}(k)$, the envelope is expanded using the inverse of the compression function given in equation 3-61. This operation is shown below in equation 3-70. Once the envelope has been expanded to cover the original range of the harmonic amplitudes, the spectral warping function is applied to transform the envelope from the linear frequency scale to the Mel scale. The warping functions are given in equations 3-52, 3-53, 3-71, and 3-72.

$$S_{sp}(k) = e^{S_{spc}(k)} \quad 0 \leq k < M \quad (3-70)$$

$$S_{sp}(\hat{k}) = S_{sp}(k) \Big|_{k=\hat{k}} \quad 0 \leq \hat{k} < \frac{M'}{2} - 1 \quad (3-71)$$

$$S_{sp}(\hat{k}) = S_{sp}(M' - \hat{k}) \quad \frac{M'}{2} \leq \hat{k} < M' \quad (3-72)$$

After the warping function has been applied, the linear predictive model is computed for the resulting Mel warped cubic spline envelope. The techniques for computing the model are identical to those presented earlier and are repeated here for clarity of notation only. The computation of the model coefficients is shown below in equations 3-73 through 3-75. Once again, the Levinson-Durbin recursion is used to solve equation 3-75 for the model coefficients, α_k [31], [32]. The solution for the model gain, G , is presented in a later section for both the voiced and unvoiced cases.

$$R_i = \frac{1}{M'} \sum_{k=0}^{\frac{M'}{2}-1} S_{sp}^2(\hat{k}) \cos\left(\frac{2\pi \hat{k} i}{M'}\right) \quad 0 \leq i \leq p \quad (3-73)$$

$$R_0 = R_0 \left(1 + \frac{1}{24576}\right) \quad (3-74)$$

$$\sum_{k=1}^p \alpha_k R_{|n-k|} = -R_n \quad 1 \leq n \leq p \quad (3-75)$$

Once the LP model coefficients have been calculated, they are converted to an alternate representation, known as line spectral pairs (LSP's) [10], [34]. Line spectral pairs are known to exhibit superior quantization properties when compared to predictor coefficients. The LSP's are obtained by decomposing the impulse response of the LP analysis filter into a difference and sum filters. These operations are shown below in equations 3-76 through 3-78.

$$B_p(z) = 1 + \alpha_1 z^{-1} + \alpha_2 z^{-2} + \dots + \alpha_p z^{-p} \quad (3-76)$$

$$V_{p+1}(z) = B_p(z) - z^{-(p+1)} B_p(z^{-1}) \quad (3-77)$$

$$Q_{p+1}(z) = B_p(z) + z^{-(p+1)} B_p(z^{-1}) \quad (3-78)$$

The LSP's, $\hat{\beta}_n$, are the roots of the difference and sum filters shown in equations 3-77 and 3-78. These coefficients can be obtained using traditional root finding techniques as well as other more efficient methods [26], [34].

3.1.6 Gain

The gain for the LP model is determined by calculating the ratio of the energy of the original and model spectra. For voiced speech, these energies are obtained by sampling the spectrum at the harmonics. For unvoiced speech, the energies are computed over the entire spectrum. Traditionally, the gain term for the LP model is obtained by matching the energy in the signal spectrum to the energy in the LP spectrum. This represents an average energy for the spectrum. We have found that for harmonic type coders, this type of gain calculation is inappropriate. The entire process is illustrated below in Figure 3-9.

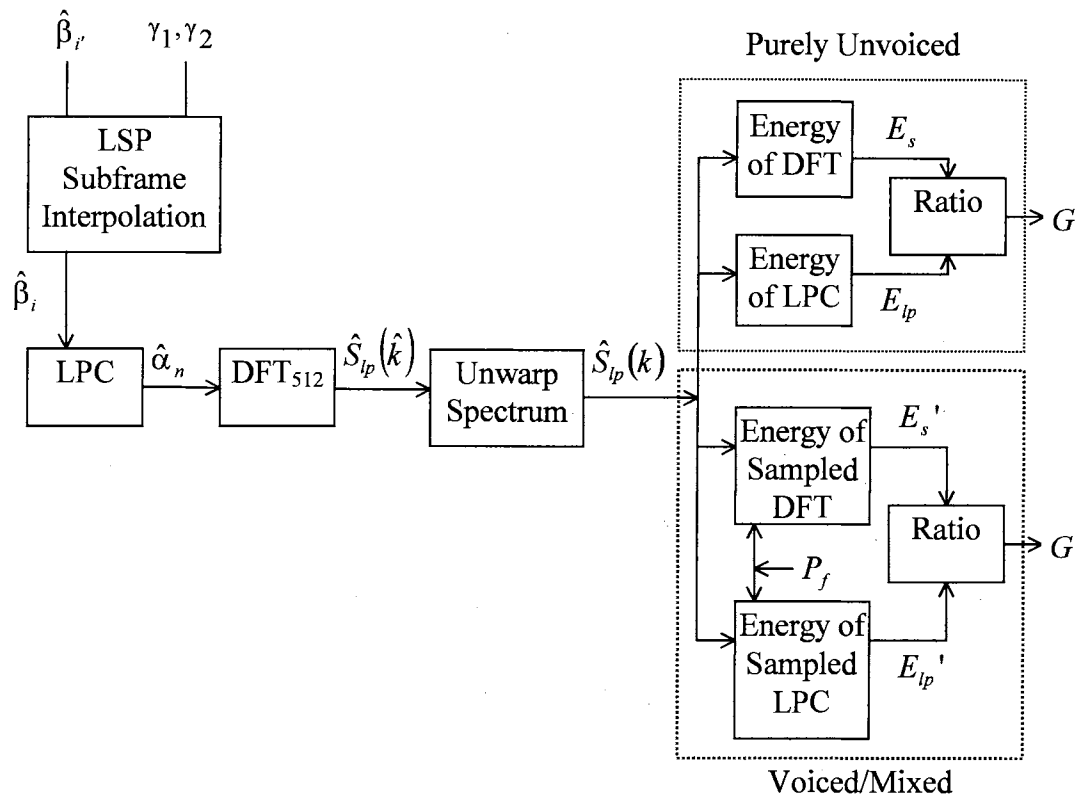


Figure 3-9. Block Diagram for Gain Calculation

The process of computing the gain begins with the conversion of the quantized line spectral pairs back to linear prediction coefficients. The gain must be calculated after the quantization of the model so that it accurately reflects the model spectrum that is being transmitted. Also, the interpolation of the LSP's across subframes must be accounted for in the gain calculation. Equation 3-79 shows the interpolation relationship for the LSP's for both superframes and subframes. The $\hat{\beta}_i(-1)$, $\hat{\beta}_i(0)$, and $\hat{\beta}_i(1)$ terms represent the LSP's for the previous frame, current frame, and future frame, and γ_1 and γ_2 correspond to the interpolation weights determined in the quantization stage.

$$\hat{\beta}_i(0) = \begin{cases} \hat{\beta}_i(0) & \text{Superframe} \\ \gamma_1 \hat{\beta}_i(-1) + \gamma_2 \hat{\beta}_i(1) & \text{Subframe} \end{cases} \quad 0 \leq i < p \quad (3-79)$$

Once the interpolation is completed, the LSP's are then converted back to prediction coefficients. Since the LSP's represent roots of the difference and sum filters given in equations 3-77 and 3-78, the LPC's are obtained by expanding these roots back out and substituting back into equation 3-76.

Following the conversion of the LSP's to LPC's, the frequency response of the LP model is obtained by computing the DFT of the model coefficients. This is shown below in equation 3-80, where M' is obtained from equation 3-53. Since the LP model is computed based on a Mel warped scale, the model spectrum must be unwarped back to the normal frequency axis. This operation is illustrated in equations 3-81 through 3-83.

$$\hat{S}_{lp}(\hat{k}) = \frac{1}{\sum_{n=0}^{M'} \hat{\beta}_n e^{-\frac{j2\pi \hat{k} n}{M'}}} \quad 0 \leq k < M' \quad (3-80)$$

$$k = \begin{cases} \hat{k} & 0 \leq \hat{k} < \frac{M'}{8} \\ \frac{M'}{8} + 1 \leq \hat{k} < \frac{M'}{2} - 1 & \frac{M'}{8} + 1 \leq \hat{k} < \frac{M'}{2} - 1 \end{cases} \quad (3-81)$$

$$\hat{S}_{lp}(k) = \hat{S}_{lp}(\hat{k}) \Big|_{\hat{k}=k} \quad 0 \leq k < \frac{M}{2} - 1 \quad (3-82)$$

$$\hat{S}_{lp}(k) = \hat{S}_{lp}(M - k) \quad \frac{M}{2} \leq k < M \quad (3-83)$$

After the LP model spectrum has been converted back to the frequency scale, the gain may be computed. For both voiced and unvoiced speech, the gain is computed as the

ratio of the energy of the spectrum of the frame, $S_s(k)$, and the energy of the LP spectrum, $\hat{S}_{lp}(k)$. For voiced speech these spectra are sampled at the location of the harmonics of the pitch for the frame. The gain computation for voiced speech is shown below in equations 3-84 through 3-86

$$E_s' = \sum_{l=1}^L S_s^2 \left(\left\lfloor \frac{LM}{P_f} + 0.5 \right\rfloor \right) \quad (3-84)$$

$$E_{lp}' = \sum_{l=1}^L \hat{S}_{lp}^2 \left(\left\lfloor \frac{LM}{P_f} + 0.5 \right\rfloor \right) \quad (3-85)$$

$$G = \sqrt{\frac{E_s'}{E_{lp}'}} \quad (3-86)$$

where L and P_f are defined by equations 3-37 and 3-45, respectively.

For unvoiced speech, the energy is computed over the entire spectrum. The gain computation for unvoiced speech is shown below in equations 3-87 through 3-89.

$$E_s = \sum_{k=1}^{\frac{M-1}{2}} S_s^2(k) \quad (3-87)$$

$$E_{lp} = \sum_{k=1}^{\frac{M-1}{2}} \hat{S}_{lp}^2(k) \quad (3-88)$$

$$G = \sqrt{\frac{E_s}{E_{lp}}} \quad (3-89)$$

3.2 Quantizer

Once the model parameters are calculated they are quantized to 2,400 bps for transmission. At this bit rate, only 72 bits are available to represent all the parameters in each 30 ms interval. The gain, pitch, and voicing decisions are coded using simple scalar quantization, while the spectral model is coded using vector quantization [34], [35]. Figure 3-10 summarizes the bit allocation and sub/superframe update scheme for each parameter.

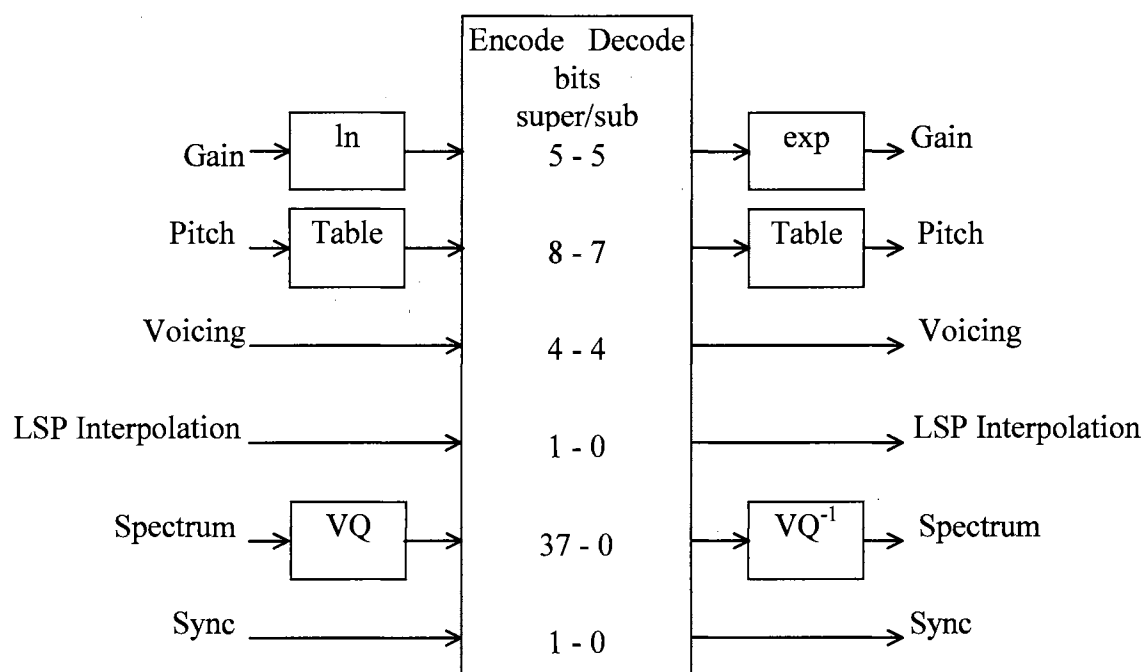


Figure 3-10. Block Diagram of Quantization and Coding

The voicing decisions are quantized as 4 bits with each bit corresponding to the voicing decision for the respective frequency band. The pitch is quantized linearly in samples from the range of 20 samples to 114 samples. For superframes, the pitch is sent as a full 8-bit value, with the subframes sending a 7-bit pitch update.

The gain, G , is logarithmically scalar quantized using equation 3-90. Equation 3-91 constrains the gain to a 5-bit range.

$$g = \lfloor 9.2 \log_{10}(1 + G) \rfloor \quad (3-90)$$

$$g' = \begin{cases} g - 14 & 14 < g \leq 46 \\ 0 & g < 14 \\ 46 & g > 46 \end{cases} \quad (3-91)$$

Prior to quantization, the LP coefficients are converted to an alternate spectral representation, known as line spectral pairs (LSP's). As mentioned previously LSP's have superior transmission and quantization properties over traditional LP coefficients.

A vector quantization (VQ) approach is used for coding the 14th order LSP model. A 37-bit, 4 way split VQ codebook is used in our current coder [35]. The 4 way split is broken down to 10, 9, 9, and 9 bits respectively. The VQ codebooks are searched for each target LSP by minimizing the squared distance between the original LSP's and the target codebook vector. The split codebooks reduce the computational complexity by allowing each codebook to represent only a small segment of the LSP spectrum and at the same time reduces the memory requirements substantially.

Once the VQ codebook entries are obtained they are transmitted on superframes only to reduce the overall bit rate. The LSP's are interpolated across subframes using a weighted linear interpolation procedure. Two candidate LSP's are computed for the subframe using the weighting shown below in equations 3-92 and 3-93, where $\hat{\beta}_i(-1)$, $\hat{\beta}_i(1)$ represent the quantized LSP's for the past and future superframes, and $\beta_i(0)$

corresponds the unquantized LSP's for the current subframe. Again, p refers to the model order.

$$\psi_0(i) = \frac{\hat{\beta}_i(-1) + 2\hat{\beta}_i(1)}{3} \quad 0 \leq i < p \quad (3-92)$$

$$\psi_1(i) = \frac{2\hat{\beta}_i(-1) + \hat{\beta}_i(1)}{3} \quad 0 \leq i < p \quad (3-93)$$

Once the interpolation candidates have been computed, an error measure is calculated between the candidates and the original unquantized LSP's. The error measure used is the weighted Euclidean distance measure shown below in equation 3-94. The weights are obtained by evaluating the spectral envelope at each LSP frequency and raising these values to a fractional power. This de-emphasizes lower energy regions, such as the formant valleys, providing a better perceptual match [35]. The weighting function is shown in equation 3-95.

$$E_\psi(k) = \sum_{k=0}^1 \sum_{i=0}^{p-1} \varepsilon(i) [\beta_i(0) - \psi_k(i)]^2 \quad (3-94)$$

$$\varepsilon(i) = \hat{S}_{sp} \left(\hat{\beta}_i(0) \frac{M' - 1}{2\pi} \right)^{\frac{1}{2}} \quad (3-95)$$

In equation 3-95, it is assumed that the LSP's lie in the range $(0, \pi]$. Once the error measures have been calculated the interpolation decision is chosen to be the candidate with the minimum error. This is shown below in equation 3-96. This interpolation decision is coded using a single bit.

$$d = \min [E_\psi(k)] \quad 0 \leq k \leq 1 \quad (3-96)$$

3.3 Synthesizer

3.3.0 Introduction

In the synthesizer, each parameter vector is recovered by reversing the encoding procedure applied in the analyzer. The vocal tract spectrum, represented as LSPs, is converted back to the coefficients of a LP model. Portions of the spectrum, which were declared unvoiced are re-synthesized by generating bandlimited noise weighted by the corresponding portion of the vocal tract spectrum. Voiced harmonics are generated as a weighted sum of harmonically related sinusoids. Between synthesis frames, the phases of corresponding harmonics must remain continuous. This is ensured by a new voiced reconstruction procedure referred to as Linear Frequency Variation (LFV) [6]. Reconstructed speech is synthesized by summing the unvoiced and voiced components. A block diagram is provided in Figure 3-11.

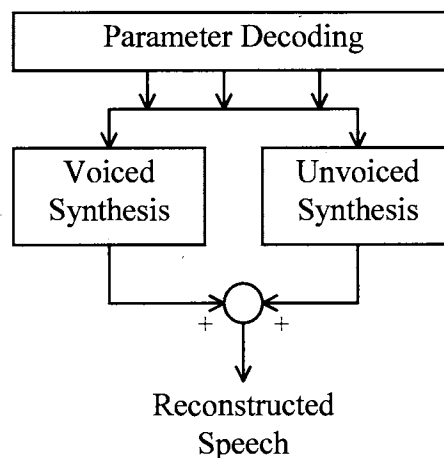


Figure 3-11. Block Diagram for Reconstructed Speech

If a band is declared unvoiced, a uniform random noise source is used to drive a bandpass filter whose passband is equal to the corresponding band. This is accomplished in the frequency domain by multiplying the voiced harmonics by zero and the unvoiced harmonics by a unity magnitude and uniform random phase. The resulting noise spectrum is then weighted by the appropriate gain. The inverse Fourier transform is computed, which results in a time domain representation for the unvoiced portion of the reconstructed speech. This is performed for each unvoiced band. The results from each band are summed producing the output reconstructed unvoiced speech for the frame.

If the current band is declared voiced, a bank of sinusoidal oscillators is used to generate a periodic signal corresponding to each harmonic in the band. These harmonics are then scaled by the appropriate harmonic amplitudes and summed. Since the pitch usually varies from frame-to-frame, voiced synthesis for this type of system becomes a problem of how to smoothly connect adjacent frames composed of sinusoids of slightly different frequencies. If the sinusoids in adjacent frames are simply "added" together, discontinuities in both frequency and phase are introduced. To overcome this problem, corresponding sinusoids in adjacent frames must be smoothed in some way so as to make the transitions smooth between frames.

This frequency smoothing must be accomplished for all corresponding harmonics between adjacent frames. The frequencies are varied linearly across the frame using LFV, and then the phase is computed at the end of the frame [6]. The frequency in the next frame is started at the ending phase of the previous frame, resulting in a frequency track with no discontinuities. This results in a smooth time domain representation for the

voiced portion. The voiced and unvoiced portions are then summed producing a frame of reconstructed speech.

It is also possible that the gain varies substantially from one frame to the next, so some amplitude smoothing is needed. This is accomplished by overlapping the reconstructed speech frames using a trapezoidal reconstruction window. The window is designed so that the sum of the overlapped windows is unity.

The following sections, Spectral Filtering, Voiced Synthesis, and Unvoiced Synthesis describe in greater detail the method used to generate synthetic speech using the sinusoidal model.

3.3.1 Spectral Filtering

This stage of the synthesizer refers to the processing of the spectral model transmitted by the analyzer. This processing includes the interpolation of the LSP's on the subframes, the computation of the gain scaled LP model, and an initial prefiltering stage to improve the perceptual quality of the spectrum. The entire spectral filtering stage is summarized in Figure 3-12.

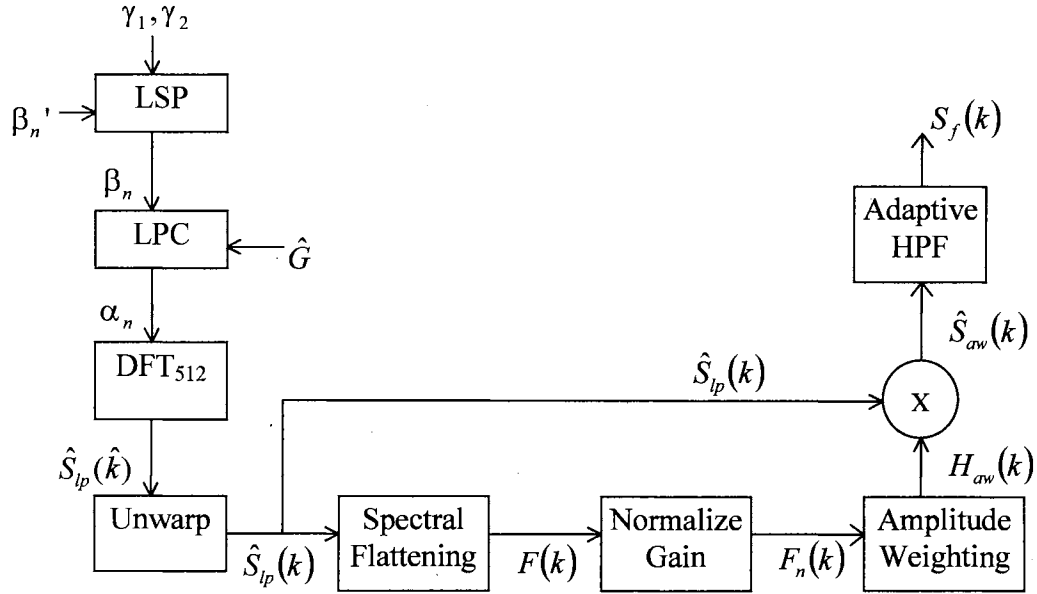


Figure 3-12. Block Diagram for Spectral Filtering

Since the LSP's are only transmitted on the superframes, an interpolation function is used to obtain spectral parameters for the subframes. The interpolation stage is identical to that observed in the gain calculation block in the analyzer. The formula is repeated below in equation 3-97 for convenience. Again, the $\hat{\beta}_{i'}(-1)$, $\hat{\beta}_{i'}(0)$, and $\hat{\beta}_{i'}(1)$ terms represent the LSP's for the previous, current, and future frames.

$$\hat{\beta}_i(0) = \begin{cases} \hat{\beta}_{i'}(0) & \text{Superframe} \\ \gamma_1 \hat{\beta}_{i'}(-1) + \gamma_2 \hat{\beta}_{i'}(1) & \text{Subframe} \end{cases} \quad 0 \leq i < p \quad (3-97)$$

Once the LSP's are obtained for the current frame they must be converted back to predictor coefficients. This is accomplished by noting that the LSP's are roots of the sum and difference filters, given in equations 3-77 and 3-78. The LP coefficients are obtained by expanding these roots and substituting into equation 3-76. Again this operation is identical to that used in the gain computation block in the analyzer.

After the predictor coefficients for the current frame have been obtained, the gain scaled spectral model for the frame is calculated. This is shown below in equation 3-98. Note that \hat{G} represents the uncoded gain value obtained by applying the inverse of equations 3-90 and 3-91 on the transmitted gain, g' . Since the predictor coefficients are initially calculated based upon the Mel scale, the spectral model must be converted back to the frequency scale. Equations 3-99 – 3-101 perform this function.

$$\hat{S}_{lp}(\hat{k}) = \frac{\hat{G}}{\sum_{n=0}^{M'} \hat{\beta}_n e^{-\frac{j2\pi\hat{k}n}{M'}}} \quad 0 \leq k < \frac{M'}{2} \quad (3-98)$$

$$k = \begin{cases} \hat{k} & 0 \leq \hat{k} < \frac{M'}{8} \\ \frac{M'}{8} \left[2^{\frac{F_s \hat{k}}{1000M}} - 1 \right] & \frac{M'}{8} + 1 \leq \hat{k} < \frac{M'}{2} - 1 \end{cases} \quad (3-99)$$

$$\hat{S}_{lp}(k) = \hat{S}_{lp}(\hat{k}) \Big|_{\hat{k}=k} \quad 0 \leq k < \frac{M'}{2} - 1 \quad (3-100)$$

$$\hat{S}_{lp}(k) = \hat{S}_{lp}(M'' - k) \quad \frac{M}{2} \leq k < M \quad (3-101)$$

The EMBE speech coder incorporates a prefiltering stage to enhance the quality of the synthetic speech signal. This prefiltering stage in the synthesizer is often referred to as postfiltering. It is known that while a speech waveform may be close to the original speech in a signal-to-noise ratio sense, it may not be close in a perceptual sense. The postfiltering stage attempts to shape the spectrum in such a way that the noise level between the formant peaks is reduced [3].

The basic operation of the postfilter is to compress the spectrum nonlinearly so that larger amplitude areas, such as formants, are relatively unaffected, while lower

amplitude regions, such as formant nulls, are significantly attenuated. Initially, a tilt corrected spectrum, $F(k)$, is computed using equation 3-102. The spectrum is flattened adaptively, based loosely on the first autocorrelation coefficient of $\hat{S}_p(k)$. The parameter that controls the level of spectral tilt compensation is shown in equation 3-104. Note that the autocorrelations are computed based upon a sampled version of $\hat{S}_p(k)$, with the samples taken at the locations of the harmonics. This is shown below in equation 3-103.

$$F(k) = \frac{\hat{S}_p(k)}{1 - \rho e^{\frac{2\pi k}{M}}} \quad 0 \leq k < \frac{M}{2} - 1 \quad (3-102)$$

$$\hat{A}(l) = \hat{S}_p(k) \bigg|_{k = \left\lfloor \frac{lM}{P_f} + 0.5 \right\rfloor} \quad 1 \leq l \leq L \quad (3-103)$$

$$\rho = \frac{\sum_{l=1}^L \hat{A}^2(k) \cos\left(\frac{l2\pi}{P_f}\right)}{\sum_{l=1}^L \hat{A}^2(l)} \quad (3-104)$$

Once the tilt-corrected spectrum is calculated, it is normalized so that it has maximum amplitude of 1.0. This is shown in equation 3-105. Following this operation the spectrum is raised to a fractional power, $\gamma = 0.3$, as shown in equation 3-106. The effect of this is to keep the formant peaks near unity amplitude, while the valleys between formants are less than unity. Raising the spectrum to a fractional power attenuates the valleys between the formants, reducing the overall noise level.

$$F_n(k) = \frac{F(k)}{\max[F(k)]} \quad 0 \leq k < \frac{M}{2} - 1 \quad (3-105)$$

$$H_{aw}(k) = F^n(k) \quad 0 \leq k < \frac{M}{2} - 1 \quad (3-106)$$

Following the calculation of the weighted normalized spectrum, $H_{aw}(k)$, the postfiltered LP model spectrum is given by equation 3-107 below.

$$S_{aw}(k) = H_{aw}(k) \hat{S}_{lp}(k) \quad 0 \leq k < \frac{M}{2} - 1 \quad (3-107)$$

An adaptive highpass filter is used to provide a slight high frequency boost and slight attenuation of lower frequencies. The adaptation of the filter is accomplished using the first normalized autocorrelation coefficient, μ , which is scaled by a constant, $K = 0.2$. The adaptive highpass filter is given in equations 3-108 and 3-109. The final LP model spectrum is denoted by $\hat{S}_f(k)$.

$$\hat{S}_f(k) = \hat{S}_{aw}(k) \left[1 - K \mu e^{-\frac{2\pi k}{M}} \right] \quad 0 \leq k < \frac{M}{2} - 1 \quad (3-108)$$

$$\mu = \frac{\sum_{k=1}^{\frac{M-1}{2}} \hat{S}_{lp}^2(k) \cos\left(\frac{k2\pi}{M}\right)}{\sum_{k=1}^{\frac{M-1}{2}} \hat{S}_{lp}^2(k)} \quad (3-109)$$

3.3.2 Unvoiced Synthesis

The unvoiced synthesis algorithm generates an unvoiced frequency spectrum $U(k)$ and then computes the inverse Fourier transform, producing a time domain sequence $u_o[n]$ that contains the unvoiced component for a given frame of reconstructed speech. A block diagram of the unvoiced synthesis is provided in Figure 3-13.

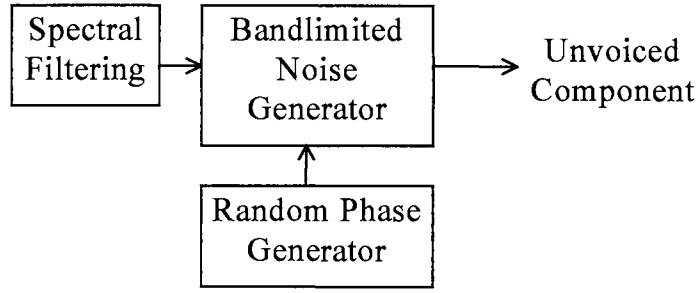


Figure 3-13. Block Diagram of Unvoiced Synthesis

Before the unvoiced frequency spectrum is generated, an initial noise spectrum $N_o(k)$ is computed by assigning a random phase component $\Theta_r(k)$ to the spectrally filtered LP spectrum $\hat{S}_f(k)$. The random phase component has a uniform distribution on the interval $[0, 2\pi)$. $N_o(k)$ is given by

$$N_o(k) = \hat{S}_f(k) e^{j\Theta_r(k)}. \quad (3-110)$$

Equation 3-110 shows the computation of the noise spectrum in terms of the magnitude and phase. Equations 3-111 and 3-112 show the noise spectrum $N_o(k)$ in terms of the real parts and imaginary parts. This noise spectrum must exhibit even symmetry for the real part and odd symmetry for the imaginary part to ensure a real time domain signal $u_o[n]$.

$$\text{Re}[N_o(k)] = \hat{S}_f(k) \cos(\theta_r(k)) \quad (3-111)$$

$$\text{Im}[N_o(k)] = \hat{S}_f(k) \sin(\theta_r(k)) \quad (3-112)$$

After the initial noise spectrum $N_o(k)$ is generated, an unvoiced spectrum $U(k)$ is generated based on the voicing decisions for each harmonic as given in equation 3-50.

For harmonics declared voiced, $U(k)$ is zeroed and the harmonics declared unvoiced are assigned as shown

$$U(k) = \begin{cases} \text{Re}[U(k)] = \text{Re}[N_o(k)] & V_l = 0 \\ \text{Im}[U(k)] = \text{Im}[N_o(k)] & V_l = 0, \\ 0 & V_l = 1 \end{cases} \quad (3-113)$$

where a harmonic l is defined on the interval described by equation 3-35.

The low and high frequency terms in the unvoiced spectrum, $U(k)$ are zeroed because the spectrum is not modeled below half of the fundamental frequency or above the last harmonic plus half of the fundamental frequency. This is described in equations 3-114 and 3-115. Again symmetry in the spectrum must be maintained.

$$U(k) = \begin{cases} \text{Re}[U(k)] = 0 & 0 \leq k \leq \text{left} \quad \text{right} \leq k \leq \frac{M}{2} \\ \text{Im}[U(k)] = 0 & 0 \leq k \leq \text{left} \quad \text{right} \leq k \leq \frac{M}{2} \end{cases} \quad (3-114)$$

$$\text{left} = \left\lceil \left(\frac{M}{2\pi} \right) \left(\frac{\omega}{2} \right) \right\rceil \quad \text{right} = \left\lceil \left(\frac{M}{2\pi} \right) (L + 0.5) \omega \right\rceil \quad (3-115)$$

The corresponding time domain signal $u[n]$ is computed using the inverse DFT as given in equation 3-116. The unvoiced synthetic speech corresponding to the current frame is overlapped with the previous frame $u_{-1}[n]$ to obtain the current reconstructed output as shown in equation 3-117 and 3-118.

$$u[n] = \frac{1}{M} \sum_{k=0}^{M-1} U(k) e^{-j \frac{2\pi k n}{M}} \quad (3-116)$$

$$u_o[n] = u_{-1}[n] w_d[n] + u[n] w_u[n] \quad (3-117)$$

$$u_{-1}[n] = u[N + n] \quad (3-118)$$

The reconstruction windows, $w_d[n]$ and $w_u[n]$, correspond to an overlapped linear taper window that is discussed in more detail in section 3.3.4.

3.3.3 Voiced Synthesis

The voiced synthesis algorithm is used to generate a time domain sequence, $v_o[n]$, that contains the current frame of reconstructed voiced speech data. A block diagram for the voiced reconstruction is given in Figure 3-14 below. A bank of sinusoidal oscillators tuned to the appropriate harmonic frequencies is used to generate the sinusoids for the voiced harmonics in a given frame. These harmonics are scaled by the corresponding harmonic amplitudes and then summed. The result is a time domain signal corresponding to the voiced components in the current frame.

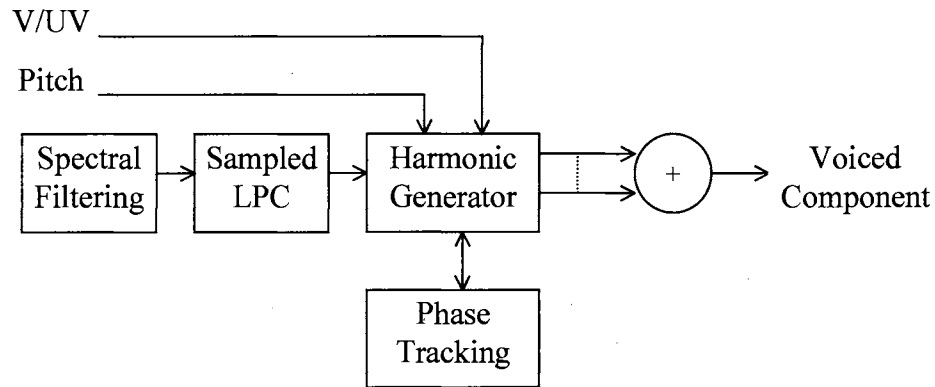


Figure 3-14. Block Diagram for Voiced Synthesis

The all voiced synthetic time domain signal is computed using equation 3-119. The sequence $v_l[n]$ is the l^{th} sinusoid with harmonic frequency $l\omega_o$, where ω_o corresponds to the fundamental frequency P_f . The upper limit on the summation is the

maximum of the number of harmonics in the previous frame, L_{-1} , and the number of harmonics in the current frame L_o .

$$v_o[n] = \sum_{l=1}^{\max\{L_{-1}, L_o\}} v_l[n] \quad 0 \leq n \leq N-1 \quad (3-119)$$

When reconstructing voiced speech using a sinusoidal model, the amplitudes must be smoothly varying from frame to frame and the frequencies and phases must be continuous at the frame boundaries. This is accomplished using the method referred to earlier as Linear Frequency Variation (LFV) [6].

There are three cases of sinusoidal reconstruction to consider. These are listed below and a development of each instance follows.

- I. A harmonic in the current frame does not have a match in the previous frame.
- II. A harmonic in the previous frame does not have a match to the current frame.
- III. A harmonic has a match from the previous frame to the current frame.

Case I occurs in two instances. One instance is that there are more harmonics in the current frame than in the previous frame, $L_o > L_{-1}$. Since the number of harmonics in each frame is not the same, the harmonics l greater than L_{-1} must be ‘born’ into the current frame. The other instance occurs when the current frame has a voiced band that is being matched to an unvoiced band in the previous frame. This means that the previous frame does not contain a matching harmonic for the current frame, and the corresponding harmonics must be ‘born’ into the current frame. Case I is generated using

$$v_l[n] = 2w_u[n]A_o(l)\cos[l\omega_o n + \theta(l)] \quad 0 \leq n \leq N-1 \quad (3-120)$$

where $\theta(l)$ is a uniformly distributed random phase generated between $[0, 2\pi)$. The harmonics are all born by using a linear taper window that is described by $w_u[n]$.

After each $v_l[n]$ has been computed, a new phase $\theta(l)$ is computed as shown in equation 3-121 in order to track the phase of the l^{th} harmonic across the frame boundary.

$$\theta(l) = l\omega_o N + \theta_{-1}(l) \quad (3-121)$$

Case II also occurs in two instances. One instance is that there are more harmonics in the previous frame than in the current frame, $L_o < L_{-1}$. Since the number of harmonics in each frame is not the same, the harmonics l greater than L_o must ‘die’ into the current frame. The other instance occurs when the current frame has an unvoiced band that is being matched to a voiced band in the previous frame. This means that the current frame does not contain a matching harmonic for the previous frame and the corresponding harmonics must ‘die’ into the current frame. Case II is generated using

$$v_l[n] = 2w_d[n]A_{-1}(l)\cos[l\omega_{-1}n + \theta(l)]. \quad 0 \leq n \leq N-1 \quad (3-122)$$

Since the harmonics are being ‘killed’ there is no reason to compute a new phase value. The harmonics are ‘killed’ by using a linear taper window that is described by $w_d[n]$.

Again, Case III occurs in two instances. If the number of harmonics in the previous frame equals the number of harmonics in the current frame, $L_o = L_{-1}$ then there is a match for every harmonic from the previous frame to the current frame. The other instance is when $L_o \leq L_{-1}$ for Case I and when $L_{-1} \leq L_o$ for Case II. If either of these are

true then there is a match for every harmonic from the previous frame to the current frame. Equation 3-123 is used to compute $v_l[n]$ for Case III

$$v_l[n] = 2(A_{-1}w_d[n] + A_o(l)w_u[n])\cos[l\omega_{-1}n + \theta(l) + \phi[n]] \quad 0 \leq n \leq N-1 \quad (3-123)$$

where $\phi[n]$ is found using equation 3-124, and a new phase value is found for $\theta(l)$ via equation 3-125.

$$\phi[n] = \frac{l(\omega_o - \omega - 1)n(n+1)}{2(N-1)} \quad 0 \leq n \leq N-1 \quad (3-124)$$

$$\theta(l) = l\omega_o N + \phi[N] + \theta_{-1}(l) \quad (3-125)$$

3.3.4 Reconstructed Output

As mentioned in the previous section, reconstructed speech must be continuous in phase and frequency and have smooth amplitude variations. This section discusses the amplitude smoothing. The block diagram in Figure 3-15 shows the application of the reconstruction window to the voiced and unvoiced components followed by summation to produce the reconstructed speech.

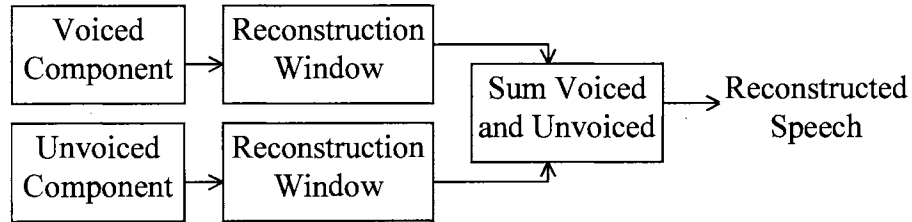


Figure 3-15. Block Diagram for Reconstructed Speech

The window used in the reconstruction process is an overlapping tapered window with the current frame centered on the overlap. This is shown in Figure 3-16. The current frame is N points in length, which corresponds to the subframe update in the analyzer.

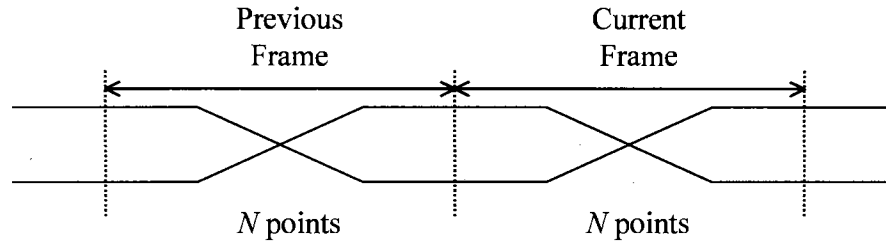


Figure 3-16. Overlapping Tapered Window

This window is described in terms of an up window $w_u[n]$ and a down window $w_d[n]$ as shown in the previous section. The equations for these two windows are given by

$$w_d[n] = \begin{cases} 1 & 0 \leq n \leq K_1 \\ 1 - \frac{n - K_1 - 1}{K_2 - K_1 - 1} & K_1 + 1 \leq n \leq K_2 \\ 0 & K_2 + 1 \leq n \leq N - 1 \end{cases} \quad (3-126)$$

and

$$w_u[n] = \begin{cases} 0 & 0 \leq n \leq K_1 \\ \frac{n - K_1 - 1}{K_2 - K_1 - 1} & K_1 + 1 \leq n \leq K_2 \\ 1 & K_2 + 1 \leq n \leq N - 1 \end{cases} \quad (3-127)$$

3.4 Results and Conclusion

As stated earlier, EMBE was developed as a candidate for the new Federal Standard at 2,400 bps. While EMBE did not win, it was competitive. Most of the problems with EMBE are related to the perceived quality, especially in the quiet and office noise environments, although EMBE has been shown to be quite robust in the harsher noise environments, such as tanks, planes, etc.

Listed below are the DAM (Diagnostic Acceptability Measure) and DRT (Dynamic Rhyme Test) scores from the May 1996 and September 1996 testing. The EMBE coder was tested in the quiet and office noise environments. The May 1996 test included improvements in the voiced reconstruction and changes in the voicing decisions (filtering). The September 1996 test included extensive work in the area of spectral modeling and perceptual filtering. Both sets of scores and the corresponding standard deviations are provided in Table 1.

The DAM is a subjective test used to determine the quality and naturalness of synthesized speech [39]. This is a comparison test so more than one vocoder is tested, where the more vocoders the more reliable the test scores. The test uses trained listeners to perform a head-to-head comparison between different vocoders.

The DRT test on the other hand is an intelligibility measure [39]. This test determines how well the synthetic speech models the initial consonant. The test is performed by generating a word list that contains word pairs according to a specified speech feature. These features are voicing, nasality, sustention, sibilation, graveness, and compactness. For example, a voicing feature contains two words where one has a voiced consonant and the other has an unvoiced consonant, such as *veal* and *feel*.

Table 1. DRT and DAM Evaluation for EMBE

May 1996				
	DRT		DAM	
	<i>Score</i>	<i>Standard Error</i>	<i>Score</i>	<i>Standard Error</i>
<i>Quiet</i>	91.1	0.86	60.3	1.2
<i>Office</i>	89.0	0.56	50.9	1.3
September 1996				
	DRT		DAM	
	<i>Score</i>	<i>Standard Error</i>	<i>Score</i>	<i>Standard Error</i>
<i>Quiet</i>	91.5	0.70	55.5	0.8
<i>Office</i>	87.7	0.66	50.7	0.9

The DRT scores from both tests appear to fall within the standard error. This suggests that no significant improvement was obtained from May 1996 to September 1996. The DAM scores in the office environment follow the same pattern as above. In the quiet environment there appears to be significant decrease in the DAM scores. This currently cannot be explained, since the September 1996 version of EMBE does sound ‘better’ to us (at Oklahoma State University) than the May version of EMBE. The rest of this dissertation is dedicated to improving the EMBE vocoder.

4 SINUSOIDAL MODEL

4.0 Introduction

The goal of this chapter is to provide a mathematical development that provides a basis for using a sinusoidal model in speech coding. The fundamental assumption of this dissertation is that analysis and synthesis of speech signals is performed using a sinusoidal model.

The sinusoidal model assumes that a frame of speech is represented by a set of frequencies, amplitudes, and phases. As noted in Chapter 2, a speech signal is in general separated into two main components: voiced and unvoiced excitation. As noted in Chapter 2, a frame of speech contains some combination of voiced and unvoiced excitation resulting in a mixed excitation frame. For voiced speech the sinusoidal model assumes that the excitation contains a harmonic structure that, when perfectly periodic, is represented by a Fourier series [3], [24]. This is the same assumption used in the *traditional* speech production model shown in Figure 2-7. The *traditional* speech production model assumed that the voiced excitation is represented by a periodic impulse train. The sinusoidal model assumes that the excitation for unvoiced speech contains an aperiodic structure that is similar to the white noise assumption of the *traditional* speech production model. By exploiting these basic properties of the speech signal a heuristic

approach is used to show that the sinusoidal model is valid for analyzing and synthesizing speech on frame-by-frame basis. This is discussed in the following sections.

4.1 Sinusoidal Model

By exploiting the properties of the Fourier transform, the validity of the sinusoidal model for analysis and synthesis is shown. Let us start by defining the Fourier transform and the properties necessary for existence. The Fourier transform of an infinite sequence $s[n]$ is a continuous set of frequencies on the range $[0, 2\pi)$, i.e., composed of all the frequencies defined on the unit circle. The forward Fourier transform of $s[n]$ is written as

$$S(e^{j\omega}) = \sum_{n=-\infty}^{\infty} s[n] e^{-j\omega n}, \quad (4-1)$$

where $s[n]$ is the infinite length input sequence and $S(e^{j\omega})$ is the corresponding Fourier transform [37]. The corresponding inverse Fourier transform of $S(e^{j\omega})$ is given by

$$s[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} S(e^{j\omega}) e^{j\omega n} d\omega. \quad (4-2)$$

Since some signals, such as speech, in general are power signals care must be taken to ensure the existence of the Fourier transform. The existence or convergence of the Fourier transform is defined by the property that

$$|S(e^{j\omega})| < \infty, \quad (4-3)$$

where $S(e^{j\omega})$ is the limit as $M \rightarrow \infty$ of the finite sum $S_M(e^{j\omega}) = \sum_{n=-M}^M s[n] e^{-j\omega n}$.

A sufficient condition for the convergence of a sequence is that it be absolutely summable [10], [37]. This condition is written as

$$\left| S(e^{j\omega}) \right| \leq \sum_{n=-\infty}^{\infty} |s[n]| < \infty \quad (4-4)$$

A sequence is made absolutely summable for analysis by splitting the infinite input sequence $s[n]$ into a set of finite length sequences (known as frames). This framing operation results in a set of concatenated finite length subsequences of $s[n]$, which are absolutely summable.

The infinite length input sequence $s[n]$ is divided into frames (finite sequences) by multiplying by a finite length, lowpass window function $w[n]$. This multiplication results in the new sequence $\tilde{s}[n]$ given by

$$\tilde{s}[n] = s[n]w[n]. \quad (4-5)$$

The sequence $\tilde{s}[n]$ is an absolutely summable sequence, and the Fourier transform converges (exists) and thus is used for analysis.

The multiplication of the input sequence $s[n]$ and window function $w[n]$ in the time-domain results in periodic convolution in the frequency domain [37]. This is defined by

$$\tilde{S}(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S(e^{j\theta}) W(e^{j(\omega-\theta)}) d\theta \quad (4-6)$$

where $W(e^{j\omega})$ is the Fourier transform of the window function $w[n]$ and is given by

$$W(e^{j\omega}) = \sum_{n=-\infty}^{\infty} w[n]e^{-j\omega n} \quad (4-7)$$

and convergence holds since the window function $w[n]$ is finite in length.

Given the above definitions and other properties of the Fourier transform, let us consider what happens if the input sequence $s[n]$ is assumed to be a constant amplitude single frequency cosine of infinite duration and written as

$$s[n] = A \cos(\omega_o n), \quad (4-8)$$

where A represents a constant amplitude, ω_o is a constant frequency, and $-\infty \leq n \leq \infty$.

Using the property that the Fourier transform of a single frequency cosine signal is an impulse located at the frequency of interest. The transform of $s[n]$ is written as

$$S(e^{j\omega}) = \pi \frac{A}{2} \delta(\omega - \omega_o), \quad (4-9)$$

where $0 \leq \omega, \omega_o < \pi$, and only the non-negative frequencies are considered. The result of the convolution shown in equation 4-6 is a scaling and shifting of the window spectrum $W(e^{j\omega})$ to the corresponding positive and negative frequencies ω_o , shown in equation 4-10, where $0 \leq \omega, \omega_o < \pi$, considering once more only the non-negative frequencies.

$$\tilde{S}(e^{j\omega}) = \frac{A}{2} W(e^{j(\omega - \omega_o)}) \quad (4-10)$$

If $\tilde{S}(e^{j\omega})$ is now sampled in the frequency domain, producing only those frequencies of interest, is the input sequence $s[n]$ recoverable? The answer is yes, if the proper sampling points are known. This sampling process is the selection of a finite set of frequencies from the Fourier transform (a sub-sampling of the Fourier transform).

Assume the frequency ω_o of the input sinusoid is known, then the spectrum is sampled at the frequency ω_o thus producing estimates of the amplitude and phase corresponding to ω_o . The sampling is equivalent to multiplying the Fourier transform

$\tilde{S}(e^{j\omega})$ by an impulse located at ω_o . The sampled Fourier transform $\tilde{S}_s(e^{j\omega})$ of the Fourier transform $\tilde{S}(e^{j\omega})$, shown in equation 4-11, is now a weighted impulse located at ω_o .

$$\tilde{S}_s(e^{j\omega}) = \frac{A}{2} W(e^{j(\omega-\omega_o)}) \delta(\omega - \omega_o) \quad (4-11)$$

As an aside, it is worth noting that the sampling did not produce the exact amplitude of the cosine defined in equation 4-8. The result is a constant A weighted by the center of the window spectrum. The center of the window spectrum is normalized by applying the constraint that the window function sum to equal 1 as given by

$$\sum_{n=-\frac{N}{2}}^{\frac{N}{2}} w[n] = 1. \quad (4-12)$$

Alternate methods to account for the weighting of the window spectrum are discussed in a later paragraph.

Given the estimate for the frequency, amplitude, and phase of the input signal $s[n]$ provided by the sampling process of equation 4-11, an estimate $\tilde{s}[n]$ for $s[n]$ is computed. This is accomplished by using the inverse Fourier transform. The inverse Fourier transform of $\tilde{S}_s(e^{j\omega})$ is found by applying the sifting property and realizing that the cosine is a real and even function. The resulting $\tilde{s}_s[n]$, shown in equation 4-13, is a weighted cosine, of finite length, with frequency equal to ω_o .

$$\tilde{s}_s[n] = \frac{AW(e^{j0})}{2\pi} e^{j\omega_o n} = \frac{AW(e^{j0})}{2\pi} \cos(\omega_o n) \quad (4-13)$$

Equation 4-13 is re-written to show that by sampling the Fourier transform at the appropriate frequencies, the original signal is recovered within a scaling factor given by

$$\tilde{s}_s[n] = G \tilde{s}[n] = G s[n] w[n], \quad (4-14)$$

where $G = \frac{AW(e^{j0})}{2\pi}$, $s[n] = \cos(\omega_o n)$, and $w[n]$ is the finite length lowpass window function.

Using the results from above, the same approach is applied to a speech sequence. First, assume that the input speech sequence is purely voiced and is represented by the sum of weighted and harmonically related cosines. The *traditional* speech production model of Figure 2-7 is used to verify this assumption [10], [12].

According to the *traditional* speech production model, either convolution in the time-domain or multiplication in the frequency-domain is used to represent continuant speech. In the time-domain an excitation sequence $e(t)$ is convolved with the impulse response $h(t)$ of the vocal tract response as shown in equation 4-15. In the frequency-domain the frequency response $E(e^{j\omega})$ of the excitation sequence $e(t)$ is multiplied by the frequency response $H(e^{j\omega})$ of the vocal tract response $h(t)$ as given in equation 4-16.

$$s(t) = e(t) * h(t) \quad (4-15)$$

$$S(e^{j\omega}) = E(e^{j\omega}) H(e^{j\omega}) \quad (4-16)$$

Now, consider the case of voiced speech. The excitation of the *traditional* speech production model is assumed to consist of an infinite duration impulse train having period T_P , where T_P corresponds to the pitch period (often referred to as the fundamental frequency). Also, assume for the time being that the vocal tract frequency response

$H(e^{j\omega})$ is a constant independent of frequency. In this case the voiced speech produced by the model of Figure 2-7 is represented by

$$s_v(t) = e(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_p). \quad (4-17)$$

The signal $s_v(t)$ in equation 4-17 is now written in terms of a Fourier series, [38], given as

$$s_v(t) = \sum_{n=-\infty}^{\infty} S_v[n] e^{j\omega_p n t}, \quad (4-18)$$

where

$$S_v[n] = \frac{1}{T_p} \int_{T_p} s_v(t) e^{-j\omega_s n t} dt. \quad (4-19)$$

Assuming the excitation is represented by the constant amplitude impulse train $s_v(t)$ as defined in equation 4-17, then by substituting into equation 4-19 and integrating over one period the result is given by

$$S_v[n] = \frac{1}{T_p} \int_{T_p} \delta(t) e^{-j\omega_s n t} dt = \frac{1}{T_p}. \quad (4-20)$$

Now, the signal $s_v(t)$ is written in a simpler form by substituting equation 4-20 into equation 4-18. The result is now written as a sum of complex exponentials given as

$$s_v(t) = \frac{1}{T_p} \sum_{n=-\infty}^{\infty} e^{j\omega_p n t} \quad (4-21)$$

and then using Euler's identity is written as the sum of weighted and harmonically related cosines given by

$$s_v(t) = \frac{1}{T_p} + \frac{2}{T_p} \sum_{n=1}^{\infty} \cos(\omega_p n t). \quad (4-22)$$

Equations 4-21 and 4-22 provide the justification for assuming that voiced speech is represented by the sum of weighted and harmonically related cosines assuming that $H(e^{j\omega})$ is a constant vocal tract frequency response independent of frequency.

Given the justification of equations 4-21 and 4-22, let the input sequence $s[n]$ be a periodic pulse train, which is represented by the sum of weighted and harmonically related cosines given as

$$s[n] = \sum_{l=1}^L A \cos(l\omega_o n), \quad (4-23)$$

where A represents the constant vocal tract response, ω_o is a constant, L is the number of sinusoids needed to represent $s[n]$, and $-\infty \leq n \leq \infty$.

The Fourier transform is a sum of harmonically related unit samples written as

$$S(e^{j\omega}) = \pi \sum_{l=1}^L \frac{A}{2} \delta(\omega - l\omega_o), \quad (4-24)$$

where ω_o represents the pitch, L is the number of sinusoids, and as before only the non-negative frequencies are considered, and $0 \leq \omega, \omega_o < \pi$.

Substituting $S(e^{j\omega})$ into equation 4-6 and performing the frequency convolution produces a sum of images of $W(e^{j\omega})$ with each image scaled and shifted in frequency by an integer multiple of the pitch ω_o and written as

$$\tilde{S}(e^{j\omega}) = \sum_{l=1}^L \frac{A}{2} W(e^{j(\omega - l\omega_o)}). \quad (4-25)$$

Now, if $\tilde{S}(e^{j\omega})$ is sampled at the appropriate frequencies then the input sequence is recovered within a scaling factor as shown previously. For voiced speech, the appropriate frequency sampling points occur at the pitch and integer multiples of the pitch (these are the harmonics). This process is given by

$$\tilde{S}_s(e^{j\omega}) = \sum_{l=1}^L \frac{A}{2} W(e^{j(\omega - l\omega_o)}) \delta(\omega - l\omega_o). \quad (4-26)$$

As before, the inverse Fourier Transform of $\tilde{S}_s(e^{j\omega})$ is found by applying the sifting property and realizing that the cosine is a real and even function. The resulting $\tilde{s}_s[n]$, shown in equation 4-27, is a weighted sum of harmonically related cosines, of finite length with frequencies equal to $l\omega_o$, where $1 \leq l \leq L$.

$$\tilde{s}_s[n] = \sum_{l=1}^L \frac{AW(e^{j0})}{2\pi} e^{jl\omega_o n} = \sum_{l=1}^L \frac{AW(e^{j0})}{2\pi} \cos(l\omega_o n) \quad (4-27)$$

Assuming the *traditional* speech production model is valid, than a voiced speech sequence is recovered within a scaling factor from a sampled version of the Fourier transform using a sinusoidal model.

If we return to our original assumptions of the *traditional* speech production model, voiced speech is synthesized from an excitation signal which is an infinite impulse train with period T_p convolved with the vocal tract response. If the vocal tract response is no longer considered to be a constant independent of frequency, the vocal tract response becomes a frequency domain weighting function $A(l)$ such that we are able to write

$$S(e^{j\omega}) = \pi \sum_{l=1}^L \frac{A(l)}{2} \delta(\omega - l\omega_o), \quad (4-28)$$

For generality, equation 4-28 is extended to include a phase term, which implies that the input signal $s[n]$ is any arbitrary sinusoid defined by

$$S(e^{j\omega}) = \pi \sum_{l=1}^L \frac{A(l)}{2} e^{j\phi(l)} \delta(\omega - l\omega_o) \quad (4-29)$$

where $e^{j\phi(l)}$ is a constant phase term. Performing the frequency convolution of equation 4-6 and applying the sampling function results in

$$\tilde{S}_s(e^{j\omega}) = \sum_{l=1}^L e^{j\phi(l)} A(l) W(e^{j0}) \delta(\omega - l\omega_o). \quad (4-30)$$

Once again, the inverse Fourier Transform of $\tilde{S}_s(e^{j\omega})$ is found by applying the sifting property and realizing that the cosine is a real and even function. The resulting $\tilde{s}_s[n]$, shown in equation 4-31, is a weighted sum of harmonically related finite length cosines with frequencies equal to $l\omega_o$, where $1 \leq l \leq L$, rotated by a constant phase term $\phi(l)$, and scaled by a frequency dependent weighting function.

$$\begin{aligned} \tilde{s}_s[n] &= \sum_{l=1}^L \frac{A(l) W(e^{j0})}{2\pi} e^{j(l\omega_o n + \phi(l))} \\ &= \sum_{l=1}^L \frac{A(l) W(e^{j0})}{2\pi} \cos(l\omega_o n + \phi(l)) \end{aligned} \quad (4-31)$$

The sampling process presented above, in equation 4-30, is contrasted to the Discrete Fourier Transform (DFT). The DFT also performs a sampling of the Fourier transform. This sampling occurs at N evenly spaced frequencies on the unit circle, where N is the length of the DFT. The DFT returns exactly N frequencies with a spacing of $2\pi/N$ as shown by

$$\tilde{S}(k) = \tilde{S}(e^{j\omega}) \Big|_{\omega = \frac{2\pi k}{N}}, \quad (4-32)$$

where $0 \leq k \leq N - 1$.

The sampling method presented in this section takes on a form similar to equation 4-32. This sampling method samples the unit circle at L evenly spaced frequencies on the unit circle, where L is the number of harmonics (sinusoids) and is determined by the pitch ω_o . The frequency spacing is given by $2\pi/L$ and is defined as

$$\tilde{S}(l) = \tilde{S}(e^{j\omega}) \Big|_{\omega = l\omega_o}, \quad (4-33)$$

where $1 \leq l \leq L$ and L is the number of harmonics. This is a more general form than equation 4-27. If the input sequence is assumed to be a sum of harmonically related sinusoids then the sampled spectrum $\tilde{S}(l)$ is in general complex and is computed from

$$\tilde{S}(l) = \tilde{S}(e^{jl\omega_o}) = \sum_{n=0}^{N-1} s[n]w[n]e^{-jl\omega_o n} \quad (4-34)$$

The inverse transform is computed by representing the complex spectrum $\tilde{S}(l)$ in terms of its magnitude and phase, shown in equation 4-35, and writing the complex exponential in terms of cosines and sines. Realizing the inverse transform must be real, the sine terms cancel leaving only cosines weighted by the magnitude of the sampled spectrum and scaled by the number of harmonics L resulting in an equation that is of the same form as equation 4-31.

$$\begin{aligned}
\tilde{s}[n] &= \frac{1}{L} \sum_{l=1}^L \tilde{S}(l) e^{jl\omega_o n} \\
&= \frac{1}{L} \sum_{l=1}^L |\tilde{S}(l)| e^{j \arg(\tilde{S}(l))} (\cos(l\omega_o n) + j \sin(l\omega_o n)) \\
&= \frac{1}{L} \sum_{l=1}^L |\tilde{S}(l)| \cos(l\omega_o n + \arg(\tilde{S}(l)))
\end{aligned} \tag{4-35}$$

The frequency sampling process described above is actually an undersampling of the DFT. Using this, we know that the energy in the original sequence and the estimated sequence is not equal. This is accounted for by multiplying by a gain function G . Thus, voiced speech is synthesized using a sinusoidal model and a scaling function G under the assumptions shown implicitly in Figure 2-7. This is written as

$$\tilde{s}[n] = \frac{G}{L} \sum_{l=1}^L |\tilde{S}(l)| \cos(l\omega_o n + \arg(\tilde{S}(l))). \tag{4-34}$$

The same arguments are developed for input speech that is considered to be unvoiced. This development is not necessary however, since unvoiced signals are generated using the sinusoidal model if we assume that the spectrum is sampled densely enough. Literature has shown that sampling with 100 Hz spacing in the frequency domain is acceptable for reconstructing noise like signals using the sinusoidal model [3]. This is equivalent to setting the pitch ω_o in equation 4-35 to 100 Hz or less.

Equation 4-36 is developed under the assumption that the excitation signal is an infinite impulse train with period T_p . Contrary to the original assumption, speech in general is a non-stationary signal. Therefore, analysis and synthesis is performed over a small time period (10 - 30 ms) where speech is considered stationary over that time period. This is a widely used assumption and leads to short time analysis and synthesis.

The two speech models discussed in Chapter 2, STC and MBE, are contrasted to equation 4-36. In the case of STC, equation 4-36 is equivalent to sampling at L unequally spaced points on the unit circle without regard to whether the frame is either voiced or unvoiced. For MBE, equation 4-36 represents how a frame (band) of voiced speech is modeled. The sampling would occur at L equally spaced points on the unit circle, equivalent to equation 4-36. If the frame (band) of speech is unvoiced then MBE uses band limited white noise in the synthesis.

4.2 Conclusion

The goal of this section was to describe the application of the sinusoidal model to a speech sequence. In the following chapter, the results of this chapter are used for developing two analysis-by-synthesis techniques based on the sinusoidal model. The first approach is in the frequency domain using only the magnitude spectrum, and the second approach is in the time domain using the magnitude spectrum and phase response.

5 SINUSOIDAL MODEL ANALYSIS-BY- SYNTHESIS

5.0 Introduction

The sinusoidal model is chosen as the topic of this dissertation because sinusoidal based vocoders have been shown to be able to produce high quality speech at low bit rates. The main disadvantage of using the sinusoidal model in developing low bit rate vocoders is the high dependence on the parameter estimation, especially the pitch. The goal of this chapter is to apply the technique of analysis-by-synthesis to the problem of parameter estimation for sinusoidal based vocoders.

In this chapter a mathematical development for two novel analysis-by-synthesis methods of parameter estimation for a vocoder, using sinusoidal excitation to synthesize high quality speech, is presented. The development is similar to methods used in Linear Prediction Analysis-By-Synthesis (LPABS) systems, such as Multi-Pulse Linear Prediction and CELP [22], [26]. The main difference between LPABS systems and the development presented in this chapter is in the method used for synthesizing speech. The synthesis technique, as previously stated, is based on the sinusoidal model.

In this dissertation a sinusoidal synthesis procedure is included in the analysis loop to determine the appropriate model parameters for the sinusoidal model. The main advantage, and the main goal of this dissertation, for including the synthesis method in

the analysis is to aid in determining the appropriate model parameters. This leads to a closed-loop analysis-by-synthesis procedure for determining the sinusoidal model parameters. By using a closed-loop approach, the parameters of the model are varied in a systematic way to produce a set of parameters that produce a synthetic signal, which matches the original signal with minimum error. This is true assuming that the model, in this case the sinusoidal model, is valid to begin with.

Assuming a frame of speech is modeled accurately using the sinusoidal model, as presented in Chapter 4, then a technique is needed to determine the appropriate set of amplitudes, frequencies, and phases (the model parameters) used to represent a frame of speech. In a manner similar to STC and MBE, the DFT is utilized in the analyzer for extracting the amplitudes, frequencies, and phases for the sinusoidal synthesis procedure. Determining these parameters is the goal of this chapter and is accomplished by developing an analysis-by-synthesis technique to improve the parameter estimation for sinusoidal based vocoders. Two novel analysis-by-synthesis methods are presented in this chapter. The first approach is developed in the frequency-domain and the second is developed in the time-domain.

The following sections provide a general overview of the current analysis and synthesis techniques being used in sinusoidal vocoders and a discussion of analysis-by-synthesis techniques. This is followed by a more specific discussion of the techniques used for the analysis and synthesis. Then a mathematical development of the two novel analysis-by-synthesis methods, frequency-domain and time-domain, is provided.

5.1 Overview

The analysis portion of the methods developed in this chapter is based on the sinusoidal model as described in Chapter 4. The sinusoidal model assumes that the input speech is represented accurately by a “set of sinusoids” having specified amplitudes, frequencies, and phases, estimated from the DFT on a frame-by-frame basis. The magnitude and phase response of the DFT are linearly sub-sampled at a specified sub-sampling period, in each frame, producing an estimate for the amplitudes, frequencies, and phases. The frequencies are computed as integer multiples of the candidate sub-sampling period, which is in contrast to STC but similar to MBE. The STC approach requires an independent frequency estimate for each of the peak amplitudes found in the DFT magnitude. By linearly sub-sampling the magnitude and phase response of the DFT, the number of parameters needed to represent a frame of speech is reduced when compared to STC. The MBE approach assumes that the peaks in the magnitude spectrum are related harmonically for voiced speech and are represented by the pitch.

The first assumption made in this development is that the amplitudes and phases for a frame of speech are known values that are determined by sub-sampling the magnitude and phase responses of the DFT at a specified sub-sampling period. The second assumption is that a set of candidate sub-sampling periods is known *a priori*. This assumption is valid for speech signals because the candidate sub-sampling periods correspond to the pitch. It is commonly known that the pitch range has a lower bound and an upper bound [10], [12]. Following these assumptions, the problem is to determine the sub-sampling period that selects the appropriate set of amplitudes and phases that produce the best match to the original input speech in a mean-squared error (MSE) sense.

In order to determine the “best” sub-sampling period, an appropriate set of candidate sub-sampling periods must be chosen. The most common sub-sampling (pitch) range for speech analysis is 70 Hz to 400 Hz (114 samples to 20 samples). Each sub-sampling candidate is used to sub-sample the DFT to obtain the corresponding amplitudes and phases. A frame of synthetic speech is generated from each set of model parameters using a sinusoidal synthesis technique. The synthesized speech, generated from each candidate sub-sampling period, is compared to the original input speech in a mean-squared error sense. The candidate sub-sampling period and the corresponding amplitudes and phases that produce the minimum mean-squared error are chosen to represent the current frame of speech.

The model parameters determined from each of the candidate sub-sampling periods are applied to the sinusoidal synthesis procedure. The result of this synthesis is compared, on a frame-by-frame basis, to the original input speech in a mean-squared error sense. This process is repeated for each candidate sub-sampling period until the model parameters producing the minimum mean-squared error have been determined.

Since the approach in this dissertation is to develop an analysis-by-synthesis routine for reconstructing high quality speech, the synthesis method plays a critical role in the design of the analysis-by-synthesis procedure. The synthesis section of this chapter discusses the common methods of reconstruction used in STC, MBE, and EMBE and details the method chosen for this dissertation.

Once the analysis and synthesis techniques are determined the analysis-by-synthesis loop is defined. Since the goal of this chapter is to develop two novel solutions to the analysis-by-synthesis problem it seems appropriate to discuss the approach used in

LPABS systems and the approach used by other sinusoidal based vocoders such as STC and MBE.

A LPABS system consists of a time varying filter, excitation signal, and perceptually based minimization and is in general closed-loop. The excitation is used to generate an estimate for the synthetic signal. This estimate is subtracted from the original signal producing an error (residual) signal. The error signal is now minimized by some method. The general approach is to use the mean-squared error criterion. While the mean-squared error criterion produces adequate performance, it is shown that a perceptual criterion works “better”. So the residual is then filtered with a perceptual weighting filter (broadening of the formant bandwidths). Now the residual is used to select another excitation sequence and a new synthetic signal is generated. The closed-loop minimization is continued until a set of parameters that produce the minimum error has been determined. This technique is discussed further in Chapter 2 regarding the CELP implementation, and more interested readers are referred to [26].

The above analysis-by-synthesis approach is a closed-loop system as most analysis-by-synthesis techniques are by nature, but STC and MBE use analysis-by-synthesis for initial mathematical development (model development) but generally estimate the parameter set in an open-loop fashion. The time-domain analysis-by-synthesis approach in STC is used to show that by selecting the peaks in the magnitude spectrum along with the corresponding frequencies and phase estimates the error between the original signal and the synthetic signal is minimized using the sinusoidal model for synthesis.

MBE uses a slightly different development than the previous LPABS and STC approaches. The first step is to estimate an initial pitch. The initial pitch is varied over small increments producing a set of pitch candidates. A set of harmonic spectral amplitudes is generated for each of the pitch candidates by sampling the magnitude spectrum. These two parameters, a refined pitch and the corresponding harmonic spectral amplitudes, are used to generate an all-voiced synthetic magnitude spectrum. The original magnitude spectrum is then compared to the synthetic magnitude spectrum in a mean-squared error sense. The pitch and the corresponding harmonic amplitudes that produce the minimum error are selected to represent the current frame of speech. Neither STC nor MBE uses the synthesis model in the analysis to aid in estimating the parameter set. This result leads to the novel developments presented in this chapter.

The two analysis-by-synthesis methods described in the following sections demonstrate that an analysis-by-synthesis approach can be used to determine a sub-sampling period that provides the appropriate amplitudes and phases to produce speech that is indistinguishable from the original when using a sinusoidal model for reconstruction. The frequency-domain analysis-by-synthesis method is developed using only the magnitude spectrum; the assumption is that no phase information is available. With this assumption, the frequency-domain analysis-by-synthesis approach is a natural application for low bit rate vocoders, at approximately 4,800 bps and lower. In contrast to the frequency-domain method, the time-domain analysis-by-synthesis method assumes that phase information is available. Since the phase information is available, the time-domain approach targets the higher bit rate sinusoidal based vocoders, approximately 12,000 bps and up.

A complete description of the analysis procedure is described in section 5.2. In section 5.3, the common reconstruction techniques are described. Section 5.4 is the development of the frequency-domain analysis-by-synthesis method of parameter estimation. In section 5.5, the time-domain analysis-by-synthesis approach is presented for parameter estimation.

5.2 Analysis

In this section, the analysis approach used for both analysis-by-synthesis developments is presented in greater detail. The analysis is conceptually a straightforward process. The frequency-domain analysis-by-synthesis approach uses the magnitude spectrum and voicing decisions to determine the appropriate amplitudes, for a given candidate sub-sampling period, to apply to the sinusoidal model. The time-domain analysis-by-synthesis approach uses the magnitude spectrum and the phase response to determine the appropriate amplitudes and phases, for a given candidate sub-sampling period, to apply to the sinusoidal model. A block diagram of only the analysis portion of the analysis-by-synthesis process is shown in Figure 5-1.

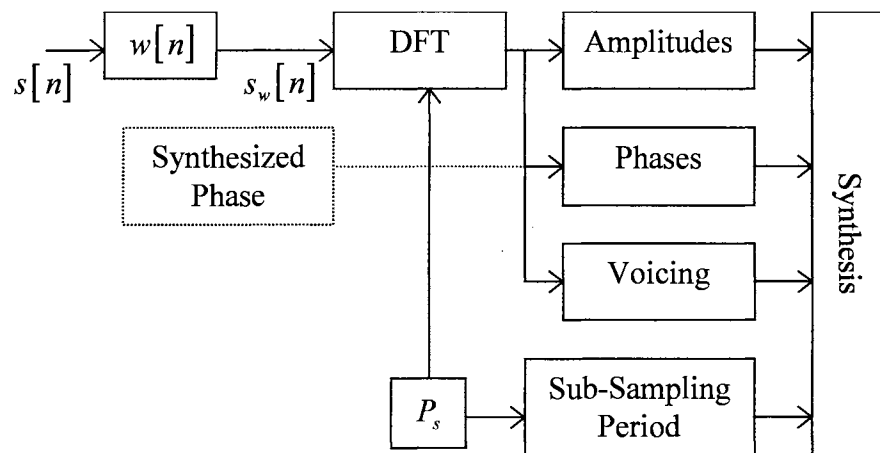


Figure 5-1. Block Diagram of Analysis Procedure

The analysis begins by separating the input speech into frames, referred to as vectors, for processing. This is accomplished by using a finite length lowpass window function $w[n]$. The new windowed speech vector is represented as

$$s_w[n] = s[n] w[n], \quad (5-1)$$

where $0 \leq n \leq N-1$, and N represents the length of the data segment (frame).

The spectrum $S_w[k]$ of the windowed signal $s_w[n]$ is then computed via the Discrete Fourier Transform (DFT) from

$$S_w[k] = \sum_{n=0}^{N-1} s_w[n] e^{\frac{j2\pi kn}{M}}, \quad (5-2)$$

where $0 \leq k \leq M-1$, and M represents the length of the DFT. The length M of the DFT is chosen to provide sufficient resolution in selecting the appropriate amplitudes and phases for a given set of candidate sub-sampling periods.

The magnitude spectrum $S_m[k]$ and phase response $S_\theta[k]$ (if necessary) are computed from the DFT using

$$S_m[k] = \sqrt{\text{Re}(S_w[k])^2 + \text{Im}(S_w[k])^2} \quad (5-3)$$

and

$$S_\theta[k] = \tan^{-1} \left(\frac{\text{Im}(S_w[k])}{\text{Re}(S_w[k])} \right). \quad (5-4)$$

The magnitude spectrum $S_m[k]$ is then sub-sampled at a rate specified by the candidate sub-sampling period P_s . If phase information is available, then the phase response $S_\theta[k]$ is also sub-sampled at the specified candidate sub-sampling period P_s . If

the phase information is not available, then it must be synthesized in the analyzer and synthesizer.

A typical range for the candidate sub-sampling periods is $20 \leq P_s \leq 114$ samples (approximately 70 Hz to 400 Hz) at an assumed input signal sampling rate of 8,000 samples per second. The discrete sub-sampling points are determined by

$$m_l = \left\lfloor \frac{M}{P_s} l + 0.5 \right\rfloor, \quad (5-5)$$

where $0 \leq m_l < \frac{M}{2}$, m_l represents the sub-sampling points in the DFT, and $1 \leq l \leq L$ with L representing the maximum number of sampling points. For example, if $M = 1024$ and $P_s = 100$ samples (80 Hz if the sampling frequency is 8,000 samples per second) then the vector m_l would equal the following

$$m_l = [10, 20, 31, \dots, 501] \quad (5-6)$$

The maximum number of amplitudes and phases being estimated for each candidate sub-sampling period is L . This number varies from one candidate sub-sampling period to another and is determined using

$$L = \alpha \frac{P_s}{2}, \quad (5-7)$$

where α is chosen to span the frequency range of interest and is generally selected to fall in the range $0.925 \leq \alpha < 1.0$.

The sub-sampled amplitudes and phases are represented by A_l and θ_l while the voicing decisions are represented by v_l . The parameters A_l and θ_l are estimated by sub-

sampling $S_m[k]$ and $S_\theta[k]$ at the points defined by equation 5-5 for a particular candidate sub-sampling period as given by

$$A_l = S_m[m_l] \quad (5-8)$$

$$\theta_l = S_\theta[m_l]. \quad (5-9)$$

If the phase information is not available, as is the case for the frequency-domain analysis-by-synthesis approach, it must be synthesized. This is accomplished by fitting a cubic spline curve to the logarithm of the spectral amplitudes that correspond to the candidate sub-sampling period. If a linear phase model is assumed, then it is necessary to compute the system phase [24]. This system phase is found using the cepstrum. As published, 44 cepstral coefficients Q are sufficient for modeling the cubic spline envelope [24]. The cepstral coefficients are computed from

$$c_m = \frac{1}{\pi} \int_0^\pi \log|H(\omega)| \cos(m\omega) d\omega \quad m = 0, 1, 2, \dots, Q \quad (5-10)$$

and the system phase is found from

$$\Phi(\omega) = -2 \sum_{m=1}^Q c_m \sin(m\omega). \quad (5-11)$$

This leads to an alternate phase representation, which is a linear phase term plus the system phase as given by

$$\theta_l = l\phi_0 + \Phi(m_l) \quad (5-12)$$

where $\phi_0 = \phi_0^{-1} + \frac{2\pi \left(\frac{1}{P_{-1}} + \frac{1}{P_0} \right) T}{2}$ and $1 \leq l \leq L$. The $-l$ implies that the parameter is

associated with the previous frame and 0 implies that parameter(s) are associated with the

current frame. A more detailed discussion of the synthetic phase model is provided in Appendix A1.

The voicing decisions are made in the frequency domain using the method detailed in Chapter 3. The sub-sampling period P_s and the corresponding amplitudes A_l and phases θ_l are applied to the synthesis model to reconstruct a frame (vector) of synthetic speech.

Two methods presented in Chapter 2 for determining the model parameters of a sinusoidal model are MBE and STC. The model parameters for MBE are determined by sampling the magnitude spectrum of a frame of speech at a specified pitch to determine the appropriate spectral amplitudes. The MBE model assumes that the amplitudes are related harmonically to the pitch, while the phase information is generated artificially in the synthesis routine. The lack of phase information for sinusoidal vocoders requires the use of a voicing decision. These voicing decisions are the heart of the MBE speech model. The magnitude spectrum for a frame of speech is divided up into a number of predefined frequency bands and each band is assigned a single voicing decision, hence the name MultiBand Excitation (MBE). The pitch parameter is determined in an open-loop fashion and is the key to the success of the MBE based vocoders. These model parameters are then applied to a sinusoidal reconstruction procedure. This approach has been shown to be capable of producing high quality speech at bit rates of 4,800 bps and up [19], [27]. One major disadvantage with the MBE implementations is the high dependence on the pitch parameter.

The model parameters for STC are determined by searching for all the peak amplitudes contained in the magnitude spectrum for a frame of speech. After the peak

amplitudes have been located, the corresponding frequencies and phases are estimated from the magnitude spectrum and the phase response. The amplitudes are not assumed to have any harmonic relationship to the pitch, although they often do for voiced speech. Since no harmonic relationship is assumed, there is no pitch estimate. The lack of a pitch estimate requires that a frequency estimate for each of the peak amplitudes be transmitted to the receiver. No voicing decisions are needed because the phase information is transmitted to the receiver. These model parameters are applied to a sinusoidal reconstruction procedure similar to that of MBE, but using a slightly different approach. The STC approach has also been shown to be capable of producing high quality speech [3]. One major disadvantage is the amount of information needed to represent a frame of speech if the standard STC model as defined above is used.

For the analysis developed in this dissertation, each input vector is analyzed for the appropriate set of amplitudes, phases, and corresponding frequencies for the sinusoidal model. These parameter estimates are then used in the sinusoidal synthesis model to generate a vector of synthetic speech. This synthetic speech vector is then passed on to the analysis-by-synthesis loop. The following section discusses the technique chosen for reconstructing synthetic speech using the sinusoidal model.

5.3 Synthesis

In this section, the synthesis approach that is used for both analysis-by-synthesis methods is presented in greater detail. The synthesis is a straightforward process but has two possible options: 1) the phase information is either generated artificially or from an assumed phase model in the synthesizer routine as in the frequency-domain analysis-by-

synthesis approach, or 2) the phase information is transmitted as in the time-domain analysis-by-synthesis approach. A general block diagram of only the synthesis portion of the analysis-by-synthesis process is shown in Figure 5-2.

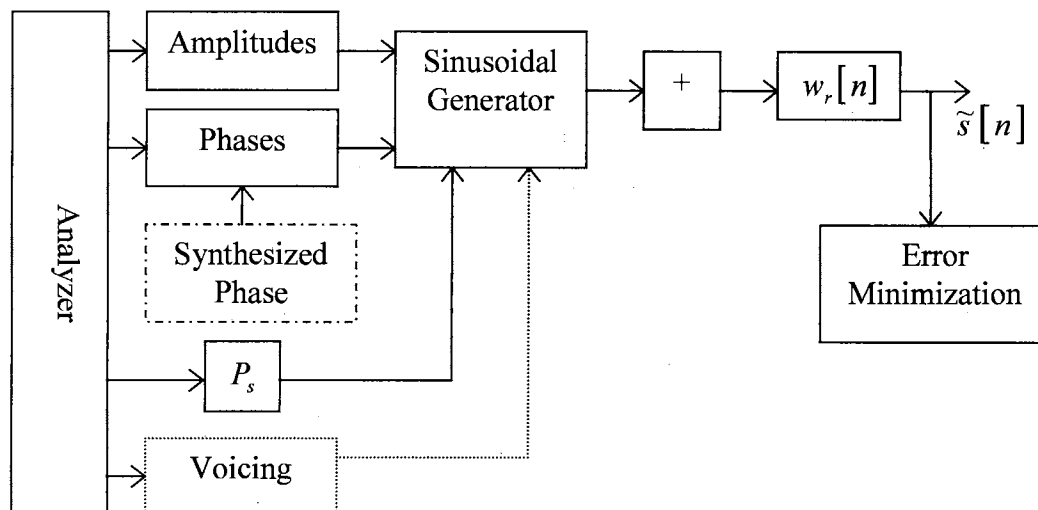


Figure 5-2. Block Diagram of Synthesis Procedure

The synthesis begins by generating a set of sinusoids that correspond to the candidate sub-sampling period P_s and its harmonics. Each of the sinusoids is weighted by the appropriate spectral amplitude A_l and rotated by the appropriate phase θ_l . These sinusoids are summed and windowed by $w_r[n]$. The window $w_r[n]$ is known as a reconstruction window and is used to smoothly connect the spectral amplitudes across frame boundaries.

The sinusoidal model used to generate the synthetic speech vector $\tilde{s}[n]$ is defined as

$$\tilde{s}[n] = \sum_{l=1}^L A_l \cos\left(\frac{2\pi m_l n}{M} + \theta_l\right) \quad (5-13)$$

where $0 \leq n \leq N - 1$, M is the length of the DFT, m_l is defined by equation 5-5, and N is the length of the reconstructed speech segment.

Equation 5-10 provides the general form for the sinusoidal synthesis model but does not provide any information about the structure of the output speech. There are a number of reconstruction methods available: cubic interpolation model, quadratic interpolation model, Linear Frequency Variation (LFV), and Overlap Addition [3], [4], [6], [24], [37]. The cubic interpolation method for the STC vocoder is discussed in Chapter 2. The cubic interpolation equations are shown in equations 2-4 and 2-5. This method of reconstruction requires a complex algorithm for matching frequencies in one frame to the frequencies in another frame. The complex frequency matching is needed to smoothly connect the sinusoids in one frame to the sinusoids of another frame to avoid frame boundary discontinuities. The complex frequency matching is performed independent of whether phase information is transmitted or generated artificially in the synthesis routine. One advantage of the STC model is that only one method of reconstruction is needed for voiced, unvoiced, or mixed speech because the phase carries voicing information. However, the synthesis frequencies must be spaced sufficiently close if the quality of the synthesized mixed or unvoiced frames is to be adequate. The disadvantage of this approach for reconstruction is the incompatibility with the concept of analysis-by-synthesis.

MBE uses a quadratic interpolation method of reconstruction. This is also presented in Chapter 2 of this dissertation. The equations for the quadratic interpolation method are shown in equations 2-8, 2-9, and 2-10. This method of reconstruction also requires a complex algorithm for matching frequencies, in order to smoothly connect, the

sinusoids in one frame to the sinusoids of another frame. The MBE model requires voicing decisions, which complicate the synthesis routine even more. The complex frequency matching is needed independent of whether the phase information is transmitted or generated artificially in the synthesis routine. The disadvantage with this model is that the quadratic model is used only for the frequencies that have been declared voiced. A separate reconstruction method is necessary for the frequencies that have been declared unvoiced. In frames that have voiced frequencies and unvoiced frequencies, the two methods of reconstruction are directly summed to produce the output speech for the frame. The unvoiced reconstruction routine uses weighted overlap addition (OLA) as discussed in chapter 2. This method of reconstruction is also incompatible with the concept of analysis-by-synthesis.

LFV is also a quadratic method of reconstruction but is only used when no phase information is transmitted [6], [7]. The phase information is not generated artificially in the synthesizer but rather it is tracked continuously across the frame boundaries forcing continuity at the pitch and the pitch harmonics. The equations used to smoothly connect the sinusoids from one frame to another are defined in section 3.2.3. The LFV method also requires voicing decisions. The frequencies declared voiced are reconstructed using LFV and the unvoiced frequencies are generated using bandpass filtered white noise. This algorithm, just like the previous two, requires a complex algorithm to match frequencies from one frame to the next. The disadvantage with this approach is the incompatibility with the proposed analysis-by-synthesis approach to parameter estimation. The LFV method requires knowledge about the pitch estimate in the future

frame in order to generate synthetic speech for the current frame, which is not feasible in the proposed analysis-by-synthesis coder.

The fourth method of reconstruction is overlap addition (OLA). The previous three methods are considered rather elegant methods for performing sinusoidal reconstruction but all three are susceptible to pitch errors. Pitch errors result in very annoying frequency chirps in the synthesized speech. The major advantage with OLA is that no complex frequency matching routine is needed to connect sinusoids across frame boundaries. Thus no future knowledge of the parameters is necessary for generating synthetic speech for the current frame. This is in contrast to the previous reconstruction methods discussed, making OLA the obvious choice for reconstructing speech for the proposed analysis-by-synthesis methods.

OLA is described in the following manner. The input speech waveform is separated into frames using a finite length lowpass overlapping window function. Each frame is analyzed independently for the sinusoidal model parameters and each frame is synthesized independently using the sinusoidal model. In order to produce correct output speech, the frames must overlap in the synthesizer. The output speech for any sample is defined in general using

$$y[n] = x[n]w[R - n] + x[n]w[2R - n] + \dots + x[n]w[NR - n]. \quad (5-14)$$

This equation assumes an overlap of N frames and R determines the amount of overlap. The result is that any sample is defined to be the sum of N numbers. This concept is illustrated in Figure 5-3 for an overlap of $N = 4$ frames, using a 240 point Hamming window, and the ratio of frame length and number of overlapping frames gives $R = 60$. The result in this case is that any output sample is the sum of 4 numbers, one

from the current frame's contribution and three from the three overlapping frame's contribution as given by

$$y[n] = x[n]w[R-n] + x[n]w[2R-n] + x[n]w[2R-n] + x[n]w[4R-n] \quad (5-15)$$

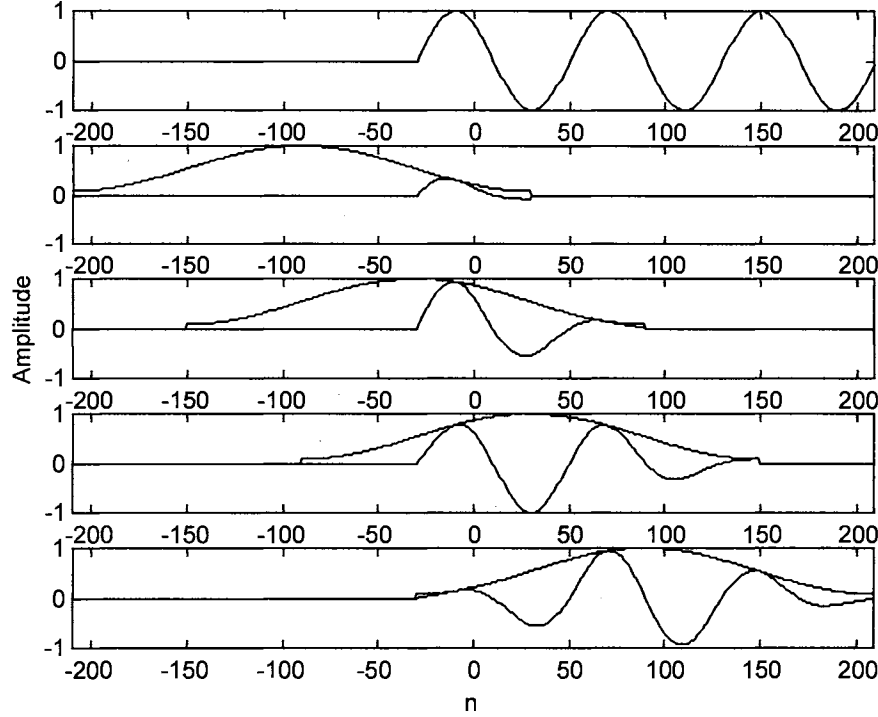


Figure 5-3. Reconstruction Using a Hamming Window

Using the concept of OLA, a formulation for the structure of the synthetic output speech is determined from equation 5-10. A frame of synthetic speech corresponding to the sinusoidal model parameters for the current analysis frame independent of the previous output is generated using equation 5-10. Then the current output $\tilde{s}[n]$ is overlapped and added to the weighted previous output samples, thus producing the actual synthetic speech estimate $\hat{s}[n]$. The amount of overlap is determined by the update rate of the analysis procedure. For example, if an overlap of 2 is being used then the full overlapped synthetic speech estimate $\hat{s}[n]$ is represented by

$$\hat{s}[n] = w_r[n]\tilde{s}^{k-1}[n] + w_r[n-N]\tilde{s}^k[n-N], \quad (5-16)$$

where $\tilde{s}^k[n]$ is the synthetic speech corresponding to the current frame's model parameters and $\tilde{s}^{k-1}[n]$ is the synthetic speech corresponding to the previous frame's model parameters.

The window $w_r[n]$ is the overlap addition reconstruction window used to smooth the overlapping speech segments. In general, the reconstruction window $w_r[n]$ is designed with the constraint that

$$w_r[n] + w_r[n-N] = 1. \quad (5-17)$$

Equation 5-17 says that the full overlap of the windows must sum to be equal to unity, whether the overlap is 2, 3, etc. A number of window types are available, including triangular, Hamming, hanning, and trapezoidal to name a few.

For the reasons previously stated in this section, the synthesis technique chosen in this dissertation is the Overlap Addition Method. Overlap addition has been shown to be capable of producing high quality synthetic speech using the sinusoidal model independent of whether the phase information is generated artificially in the synthesizer or transmitted to the synthesizer [3], [24]. The next section describes the combination of the analysis and synthesis procedures into an analysis-by-synthesis approach to determine the model parameters for the sinusoidal model.

5.4 Frequency-Domain Analysis-By-Synthesis Sinusoidal Model

5.4.0 Introduction

This section describes the development of a frequency-domain analysis-by-synthesis method. The analysis and synthesis techniques described in sections 5.2 and 5.3 are combined to form a closed-loop analysis to estimate the model parameters using a mean-squared error approach. The frequency-domain approach assumes that no phase information is transmitted so the phase information is not available. A block diagram for the frequency-domain analysis-by-synthesis approach is shown in Figure 5-4.

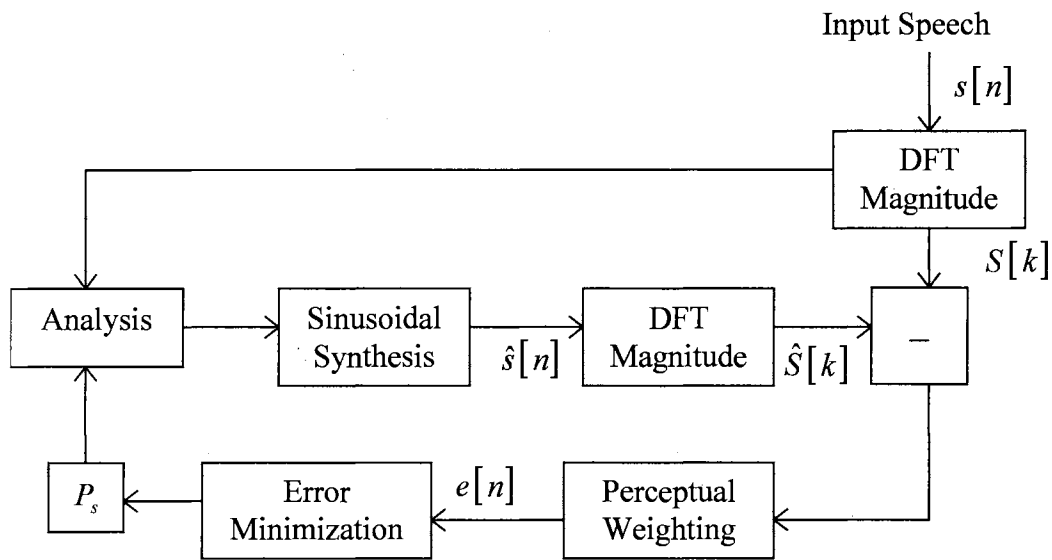


Figure 5-4. Frequency-Domain Analysis-By-Synthesis Using Sinusoidal Model

The parameters needed to synthesize speech using the sinusoidal model are pitch, voicing, synthesized phase, and spectral amplitudes. This parameter estimation is accomplished by first computing the magnitude spectrum of the input vector. Then a candidate sub-sampling period P_s is selected to represent the current analysis frame. Using the technique described in section 5.2 the amplitudes are estimated as a function of

the candidate sub-sampling period P_s . The spectral amplitudes are determined by sub-sampling the magnitude spectrum. Since the assumption is that no phase information is available in the analyzer it must be generated in the synthesizer. The method for generating the synthetic phase is discussed in section 5.2 and presented in greater detail in Appendix A2.

The next two sections provide the mathematical development for both the frequency-domain and time-domain analysis-by-synthesis methods.

5.4.1 Frequency-Domain Analysis-By-Synthesis

Using the parameter estimates presented above, the sinusoidal model is used to generate a frame of synthetic speech. The synthetic speech vector is subtracted from the input speech vector, and a mean-squared error value (MSE) is computed. The MSE is used to select another candidate sub-sampling period, which selects an alternate set of spectral amplitudes. This process is repeated until a MSE value is computed for each candidate sub-sampling period. The sub-sampling period that corresponds to the set of model parameters having the minimum MSE is chosen to represent the current frame of input speech.

This procedure is written mathematically using the common notation for the mean-squared error as

$$E = \frac{1}{N} \sum_{n=0}^{N-1} e^2[n] \quad (5-18)$$

where $e[n] = s_{w_r}[n] - \hat{s}[n]$, $s_{w_r}[n]$ is the appropriately windowed input signal, and $\hat{s}[n]$ is given by equation 5-16. The input signal has the reconstruction window $w_r[n]$ applied so

that the comparison between the input signal $s[n]$ and the synthetic signal $\hat{s}[n]$ is approximately a one-to-one matching. The total error E , after substituting equation 5-16 into equation 5-18 and substituting in the reconstruction window $w_r[n]$, is given as

$$E = \frac{1}{N} \sum_{n=0}^{N-1} \left(s[n]w_r[n] - \left(\tilde{s}^{k-1}[n]w_r[n] + \tilde{s}^k[n]w_r[n-N] \right) \right)^2 \quad (5-19)$$

The main problem in this approach is the fact that by generating the phase information artificially, the time alignment between the original input signal and the synthetic signal is lost. This problem is countered by performing the analysis-by-synthesis in the frequency-domain using the magnitude spectrum and neglecting the phase response.

The obvious problem with the frequency-domain approach is that equation 5-19 would involve performing convolutions. Initially this problem is viewed as an extremely complex problem to solve because of the number of time varying parameters and the dependence on knowing information about frequency-domain parameters using a time-domain synthesis. This problem is simplified by redefining equation 5-16, making a number of assumptions about the generation of the synthetic signal, and performing the analysis-by-synthesis in the frequency-domain. This problem is redefined using knowledge about the sub-sampling process. We know that the energy in the synthetic signal is not equal to the energy in the original signal as a result of the sub-sampling process. The problem is now defined by rewriting equation 5-16 in terms of a gain g multiplied by the weighted sum of sinusoids as given by

$$\hat{s}[n] = w_r[n]g^{k-1}\tilde{s}^{k-1}[n] + w_r[n-N]g^k\tilde{s}^k[n-N]. \quad (5-20)$$

Equation 5-18 is now rewritten by substituting equation 5-20 producing the new total error E as defined by

$$E = \frac{1}{N} \sum_{n=0}^{N-1} \left(s[n]w_r[n] - \left(g^{k-1}\tilde{s}^{k-1}[n]w_r[n] + g^k\tilde{s}^k[n]w_r[n-N] \right) \right)^2. \quad (5-21)$$

This equation does not lend itself to an easy minimum solution in either the time-domain or frequency-domain. The time-domain approach still has problems with misalignment because of the lack of phase information. The frequency-domain approach is simplified by making some reasonable assumptions. The first assumption is to perform the analysis before the reconstruction window is applied. The second is that the overlap is not included in the analysis of the current frame. The total error E is now rewritten using equation 5-20 and per the assumptions as

$$E = \frac{1}{N} \sum_{n=0}^{N-1} (s[n] - g\tilde{s}[n])^2 \quad (5-22)$$

where the frame designation k has been dropped.

Note that the initial target is being defined without the overlap from the previous frame. This seems to be a correct approach since the idea is to match the frequency response of the current frame of synthetic speech with that of the input frequency response. By including the overlap, the frequency response of the synthetic speech is distorted because of the effect of the reconstruction window. The current frame of synthetic speech is defined as a gain g multiplied by the synthetic speech generated using equation 5-10. The addition of the gain term seems appropriate because the magnitude spectrum of the synthetic speech has less energy than the magnitude spectrum of the input speech. By undersampling the magnitude spectrum, quantization error is

introduced into the reconstructed speech; the energy in the reconstructed speech is not equal to the energy of the original input speech as shown in Chapter 4.

Equation 5-22 is written to reflect the frequency-domain approach by computing the magnitude spectrum of $s[n]$ and $g\tilde{s}[n]$ which is defined by

$$E = \frac{1}{M} \sum_{k=0}^{M-1} (S(k) - G\tilde{S}(k))^2 \quad (5-23)$$

where the magnitude spectra are found using the DFT, M is the length of the DFT, and it is assumed that proper windowing has been applied.

The total error E in equation 5-23 is the term that we want to minimize. The search procedure consists of finding the sub-sampling period P_s that produces a set of parameters that produces synthetic speech that best matches the input speech in a weighted least square error sense as shown in equation 5-23.

First let us expand equation 5-23 by computing the square and substituting J for E ; the reason becomes clear later. The total error J computed for each of the candidate sub-sampling periods P_s is given by

$$J^i = \frac{1}{M} \sum_{k=0}^{M-1} S^2(k) - \frac{2}{M} \sum_{k=0}^{M-1} G^i S(k) \tilde{S}(k)^i + \frac{1}{M} \sum_{k=0}^{M-1} G^{2i} \tilde{S}^2(k)^i \quad (5-24)$$

where the index i selects the i^{th} candidate sub-sampling period P_s^i and its corresponding spectral amplitudes selected by sub-sampling the magnitude spectrum. The total error J^i is associated with the current frame's parameters.

As with any minimum error scheme, defining the appropriate match criterion is key to the success of the minimization process. Initially we want to match the original input signal to the corresponding synthetic signal, and in this case we are trying to match

the magnitude spectra. The target signal to be matched is defined to be $E^{(0)}(k)$ which is given by

$$E^{(0)}(k) = S(k). \quad (5-25)$$

The substitution of equation 5-25 into 5-24 leads to the new total error term defined as

$$J^i = \frac{1}{M} \sum_{k=0}^{M-1} E^{(0)2}(k) - \frac{2}{M} \sum_{k=0}^{M-1} G^i E^{(0)}(k) \tilde{S}(k)^i + \frac{1}{M} \sum_{k=0}^{M-1} G^{2i} \tilde{S}^2(k)^i. \quad (5-26)$$

The total error J^i is still dependent on the gain term G^i and the candidate sub-sampling period P_s^i . This is still a complex problem that calls for solving for G^i and P_s^i simultaneously. An alternate approach is to solve for the two parameters sequentially. The sequential approach is as follows: solve for the optimum G^i using equation 5-26, then solve for P_s^i given the optimum gain G^i .

The gain is found by computing the partial derivative of J^i with respect to the gain G^i , and setting the result equal to zero and solving. This is given in the following two equations.

$$\frac{\partial J^i}{\partial G^i} = -2 \sum_{k=0}^{M-1} E^{(0)}(k) \tilde{S}(k)^i + 2 \sum_{k=0}^{M-1} G^i \tilde{S}^2(k)^i = 0 \quad (5-27)$$

$$G^i = \frac{\sum_{k=0}^{M-1} E^{(0)}(k) \tilde{S}(k)^i}{\sum_{k=0}^{M-1} \tilde{S}^2(k)^i} \quad (5-28)$$

Equation 5-28 is the normal form of the cross-correlation. In order to find the optimal minimum MSE sub-sampling period equation 5-26 is set equal to zero as shown by

$$J^i = \sum_{k=0}^{M-1} E^{(0)2}(k) - 2 \sum_{k=0}^{M-1} G^i E^{(0)}(k) \tilde{S}(k)^i + \sum_{k=0}^{M-1} G^{2i} \tilde{S}^2(k)^i = 0. \quad (5-29)$$

The target $E^{(0)}$ is not a function of the index i so equation 5-29 is rewritten by moving the target energy term to the left side producing

$$\sum_{k=0}^{M-1} E^{(0)2}(k) \geq 2 \sum_{k=0}^{M-1} G^i E^{(0)}(k) \tilde{S}(k)^i - \sum_{k=0}^{M-1} G^{2i} \tilde{S}^2(k)^i \quad (5-30)$$

This inequality is motivated in the following way. The term on the left side of the inequality is the autocorrelation of the target vector. This value represents the best possible match between the target vector $E^{(0)}$ and the synthesized magnitude spectrum $G^i \tilde{S}(k)^i$. This would suggest that we would want to maximize the term on the right side of the inequality. This is fine if G^i is a positive value, but the gain G^i is a quantity that is either positive or negative. Since G^i can be negative, there is a possibility that the autocorrelation of the target $E^{(0)}$ is equal to a negative value. This problem is easily solved by the fact that the only way G^i is going to be negative is if the quantity $\tilde{S}(k)^i E^{(0)}$ results in a negative value. If this happens then the term on the right side of equation 5-30 has a positive result, since $G^{(i)}$ and $\tilde{S}(k)^i E^{(0)}$ are both negative. The right side of equation 5-30 is largest when the synthetic speech vector $\tilde{S}(k)^i$ approaches the target vector $E^{(0)}$. This suggests that the optimum minimum MSE is determined by maximizing the quantity $2 \sum_{k=0}^{M-1} G^i \tilde{S}(k)^i E^{(0)} - \sum_{k=0}^{M-1} G^{2i} \tilde{S}^2(k)^i$ as shown by

$$M_s^i = 2 \sum_{k=0}^{M-1} G^i \tilde{S}(k)^i E^{(0)} - \sum_{k=0}^{M-1} G^{2i} \tilde{S}^2(k)^i. \quad (5-31)$$

Equation 5-31 is referred to as the *match score* M_s^i for the current set of model parameters and is rewritten in a more compact form by substituting the optimal gain G^i from equation 5-28 into equation 5-31. The result is the following match score

$$M_s^i = \sum_{k=0}^{M-1} G^i \tilde{S}(k)^i E^{(0)} = \frac{\left(\sum_{k=0}^{M-1} \tilde{S}(k)^i E^{(0)} \right)^2}{\sum_{k=0}^{M-1} \tilde{S}^2(k)^i}. \quad (5-32)$$

This is the squared cross-correlation of the target vector and the synthesized magnitude spectrum normalized by the energy in the synthetic magnitude spectrum corresponding to index i , which directly relates to the optimum sub-sampling period P_s^i .

In summary, a set of candidate sub-sampling periods is selected to represent the current analysis frame. This vector is denoted as P_s^i and typically ranges from 20 samples to 114 samples for speech signals. For each value of P_s^i , a gain G^i and a match score M_s^i are computed as shown in equations 5-28 and 5-32. Since the match score is a maximizing function, the sub-sampling period corresponding to the largest match score is selected to represent the current analysis frame along with the corresponding gain and amplitudes. The following paragraphs discuss the results of the frequency-domain analysis-by-synthesis method derived above.

For ease of development, clarity, and without loss of generality the development of the previous equations are written in terms of vector notation. This is acceptable since it is equivalent to dividing the input data into frames. The total error now becomes

$$\mathbf{E} = \|\mathbf{e}\|^2 = \mathbf{e}_1^2 + \mathbf{e}_2^2 + \dots + \mathbf{e}_N^2 \quad (5-33)$$

As above, the match criterion is defined as the current frame's synthetic speech scaled by a gain \mathbf{g} . The match criterion vector for the current synthesis frame is given by

$$\hat{\mathbf{s}} = \mathbf{g}\tilde{\mathbf{s}} \quad (5-34)$$

Since this is a frequency-domain approach, equations 5-33 and 5-34 must be rewritten in terms of the magnitude spectrum. The general form for computing the minimum total error, using an alternate notation, is rewritten as

$$\mathbf{J} = \|\mathbf{E}\|^2 = \mathbf{E}_1^2 + \mathbf{E}_2^2 + \dots + \mathbf{E}_N^2. \quad (5-35)$$

The new match criterion equation defined in terms of the magnitude spectrum is defined by

$$\hat{\mathbf{S}} = \mathbf{G}\tilde{\mathbf{S}}, \quad (5-36)$$

where $\hat{\mathbf{S}}$ and $\tilde{\mathbf{S}}$ are computed using equation 5-2, and the magnitude spectrum are found using equation 5-3 of the input speech signal and the synthetic speech signal, respectively.

The common form for the error vector \mathbf{J} is a perceptually weighted difference between the magnitude spectrum of the input speech vector and the magnitude spectrum of the synthetic speech vector defined as

$$\mathbf{E}^i = \mathbf{W}(\mathbf{S} - \hat{\mathbf{S}}^i), \quad (5-37)$$

where \mathbf{W} is a lower triangular matrix that represents the impulse response of the perceptual weighting filter [20], [26]. The perceptual weighting is included without loss of generality or clarity as is discussed in a later section. The index i determines the synthetic speech vector that corresponds to a given candidate sub-sampling period \mathbf{P}_s and the associated voicing decisions, phases, and spectral amplitudes.

From equation 5-37 a target vector is defined by $\mathbf{E}^{(0)} = \mathbf{W}\mathbf{S}$. Now by substituting equation 5-36 into equation 5-37 and rewriting, a new error vector is specified in terms of the target vector as

$$\mathbf{E}^i = \mathbf{E}^{(0)} - \mathbf{G}^i \tilde{\mathbf{S}}^i. \quad (5-38)$$

Substituting equation 5-38 into equation 5-35 leads to the following error metric.

$$\mathbf{J}^i = \|\mathbf{E}^i\|^2 = \mathbf{E}^{(0)\text{T}} \mathbf{E}^{(0)} - 2\mathbf{G}^i \tilde{\mathbf{S}}^{i\text{T}} \mathbf{E}^{(0)} + \mathbf{G}^{i2} \tilde{\mathbf{S}}^{i\text{T}} \tilde{\mathbf{S}}^i \quad (5-39)$$

\mathbf{J} is the total squared error sum corresponding to the candidate sub-sampling period vector \mathbf{P}_s^i , and \mathbf{T} is the transpose of the vector. Since \mathbf{J} is a function of both \mathbf{G} and i then an optimal \mathbf{G} is found for a given index i . This is accomplished by computing the partial derivative of \mathbf{J} with respect to the gain \mathbf{G} and then setting the derivative equal to zero. This is computationally shown as

$$\frac{\partial \mathbf{J}^i}{\partial \mathbf{G}^i} = -2\tilde{\mathbf{S}}^{i\text{T}} \mathbf{E}^{(0)} + 2\mathbf{G}^i \tilde{\mathbf{S}}^{i\text{T}} \tilde{\mathbf{S}}^i = 0. \quad (5-40)$$

This equation is solved to find an optimal gain \mathbf{G} , in the minimum mean-squared error sense. The optimal gain \mathbf{G} is found from computing the normalized cross-correlation between the target vector $\mathbf{E}^{(0)}$ and the synthetic speech vector $\tilde{\mathbf{S}}^i$ corresponding to index i as shown by

$$\mathbf{G}^i = \frac{\tilde{\mathbf{S}}^{i\text{T}} \mathbf{E}^{(0)}}{\tilde{\mathbf{S}}^{i\text{T}} \tilde{\mathbf{S}}^i}. \quad (5-41)$$

To determine the optimal minimum MSE sub-sampling period \mathbf{P}_s^i , equation 5-39 is set equal to zero as shown in the equation below.

$$\mathbf{J}^i = \mathbf{E}^{(0)\text{T}} \mathbf{E}^{(0)} - 2\mathbf{G}^i \tilde{\mathbf{S}}^{i\text{T}} \mathbf{E}^{(0)} + \mathbf{G}^{i2} \tilde{\mathbf{S}}^{i\text{T}} \tilde{\mathbf{S}}^i = 0 \quad (5-42)$$

Since the target vector $\mathbf{E}^{(0)}$ is not a function of index i it is moved to the left side producing the following equation.

$$\mathbf{E}^{(0)\top} \mathbf{E}^{(0)} \geq 2\mathbf{G}^i \tilde{\mathbf{S}}^{i\top} \mathbf{E}^{(0)} - \mathbf{G}^{i^2} \tilde{\mathbf{S}}^{i\top} \tilde{\mathbf{S}}^i \quad (5-43)$$

This inequality is motivated in the following way. The term on the left side of the inequality is the autocorrelation (energy) of the target vector. This value represents the best possible match between the target vector $\mathbf{E}^{(0)}$ and the synthesized speech $\mathbf{G}^i \tilde{\mathbf{S}}^i$. This would suggest that we would want to maximize the term on the right side of the inequality. This is fine if \mathbf{G}^i is a positive value, but the gain \mathbf{G}^i is a quantity that is either positive or negative. Since \mathbf{G}^i can be negative, there is a possibility that the autocorrelation of $\mathbf{E}^{(0)\top}$ is equal to a negative value. This problem is easily solved by the fact that the only way \mathbf{G}^i is negative is if the quantity $\tilde{\mathbf{S}}^{i\top} \mathbf{E}^{(0)}$ results in a negative value. If this happens then the term on the right side of equation 5-43 has a positive result, since \mathbf{G}^i and $\tilde{\mathbf{S}}^{i\top} \mathbf{E}^{(0)}$ are both negative. The right side of equation 5-43 is largest when the synthetic speech vector $\tilde{\mathbf{S}}^i$ approaches the target vector $\mathbf{E}^{(0)}$. This suggests that the optimum minimum mean-squared error index i is found by maximizing the quantity $2\mathbf{G}^i \tilde{\mathbf{S}}^{i\top} \mathbf{E}^{(0)} - \mathbf{G}^{i^2} \tilde{\mathbf{S}}^{i\top} \tilde{\mathbf{S}}^i$, as given by

$$\mathbf{M}^i = 2\mathbf{G}^i \tilde{\mathbf{S}}^{i\top} \mathbf{E}^{(0)} - \mathbf{G}^{i^2} \tilde{\mathbf{S}}^{i\top} \tilde{\mathbf{S}}^i. \quad (5-44)$$

Equation 5-44 is referred to as the match score for the current set of model parameters and is rewritten in a more compact form by substituting the optimal gain \mathbf{G} from equation 5-41 into equation 5-44. The result is the following expression for the match score

$$\mathbf{M}^i = \mathbf{G}^i \tilde{\mathbf{S}}^{i^T} \mathbf{E}^{(0)} = \frac{(\tilde{\mathbf{S}}^{i^T} \mathbf{E}^{(0)})^2}{\tilde{\mathbf{S}}^{i^T} \tilde{\mathbf{S}}^i}. \quad (5-45)$$

This is the squared cross-correlation of the target vector and the synthesized speech normalized by the energy in the synthetic speech vector corresponding to index i .

In summary, a set of candidate sub-sampling periods is selected to represent the current analysis frame. This vector is denoted as \mathbf{P}_s^i and typically ranges from 20 samples to 114 samples for speech signals. For each value of \mathbf{P}_s^i , a gain \mathbf{G}^i and a match score \mathbf{M}^i are computed as shown in equations 5-39 and 5-43. Since the match score is a maximizing function, the sub-sampling period corresponding to the largest match score is selected to represent the current analysis frame along with the corresponding gain and amplitudes. The following paragraphs discuss the simulation of the frequency-domain analysis-by-synthesis method derived above.

5.4.2 Simulation

5.4.2.0 Introduction

This section describes the simulation of the frequency-domain analysis-by-synthesis method derived in this section. The simulation is written in the 'C' programming language on a Sun Sparc Workstation Ultra 170. The idea here is to prove the concept of the frequency-domain analysis-by-synthesis approach so complexity is considered to be of secondary importance. While this is a simulation it is worth noting that since no phase information is necessary to perform the analysis, this simulation is naturally targeted towards low bit rates.

The input signals used in this simulation are quantized using 16 bits and sampled at 8,000 samples per second. The input signal is windowed using a 240 (30ms) point square-root of Hamming window. This windowed signal represents a frame of speech. This window has less smearing of the main lobe and higher side lobes as compared to a regular Hamming but has more smearing of the main lobe and greater attenuation of the side lobes as compared to a rectangular window. The analysis window is updated by shifting across in 7.5ms intervals (60 samples). Assuming the center of the window is the time reference, this update structure results in an overlap of 90 samples in the past and 90 samples in the future.

The magnitude spectrum of the input signal is computed using the DFT. As stated in section 5.2, the length of the DFT is chosen to provide appropriate resolution for selection of the spectral amplitudes. The DFT length chosen is $M = 16,384$. This length provides a resolution of approximately 0.5 Hz for 8 KHz sampled signals. This magnitude spectrum is sub-sampled producing a set of spectral amplitudes for a given candidate sub-sampling period.

The phase information is computed as presented in section 5.2. A cubic spline is fitted to the logarithm of the spectral amplitudes for each of the corresponding candidate sub-sampling periods. A set of cepstral coefficients is found for each of the corresponding candidate sub-sampling periods that provide a good fit to the cubic spline envelope. These cepstral coefficients are then used to compute the system phase for the sinusoidal model. This phase parameter is only computed at the sample points.

The range for the sub-sampling period P_s is selected to be 20 to 114 samples. While the sub-sampling period is defined on a finite range the number of possible sub-

sampling candidates is refined further. The process is divided into two stages. The first stage performs the analysis-by-synthesis on only the integer sub-sampling periods. The second stage is used to perform a refinement by searching ± 2 samples around the integer sub-sampling candidate producing the highest match score, in the stage one analysis, in 0.2 sample increments.

After obtaining the spectral amplitudes for each of the corresponding candidate sub-sampling periods, a set of voicing decisions are computed. The voicing decisions determine which spectral amplitudes in a given frame are to be generated as either voiced or unvoiced. The frequency-domain approach developed in EMBE is used to determine the voicing decisions. This method is detailed in Chapter 3.

For each of the candidate sub-sampling periods a synthetic signal is generated using the sinusoidal model with OLA. The synthetic signals are compared in a least mean-squared error sense to the original input signal. An optimum gain and a match score are found for each of the corresponding synthetic signals where the parameter set producing the highest match score is chosen to represent the current frame.

The following paragraphs discuss in more detail the signals used in testing the concept of frequency-domain analysis-by-synthesis, the sub-sampling process, the analysis-by-synthesis loop, the match scores, and the resulting sub-sampling contour.

5.4.2.1 Test Signals

Three signals are used to test the frequency-domain analysis-by-synthesis process. The first signal shown in Figure 5-5 is a constant tone that is generated from a weighted sum of sinusoids where the fundamental frequency is equal to approximately 126 Hz. This signal is used to test the response to an all voiced speech signal.

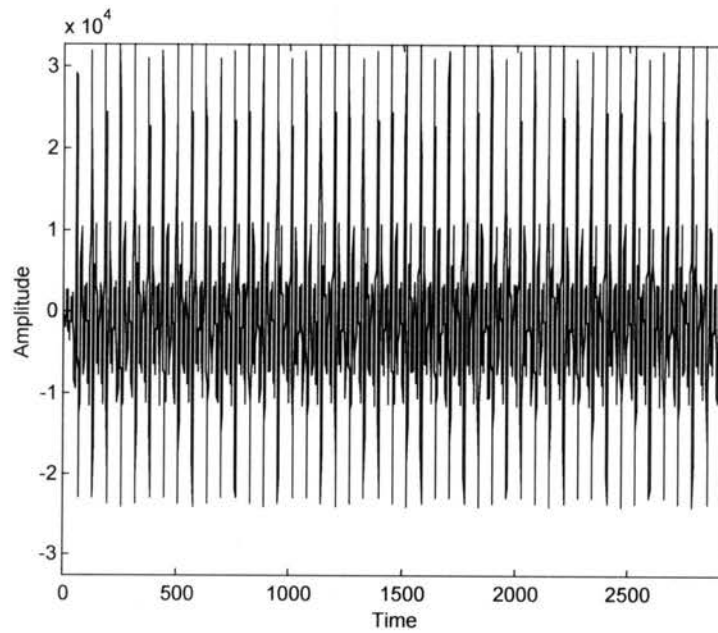


Figure 5-5. All Voiced Signal

The second signal shown in Figure 5-6 is bandlimited white noise that is also generated from a sum of weighted sinusoids with no periodic structure (the frequencies are selected arbitrarily). This signal is used to test the response to an all unvoiced signal.

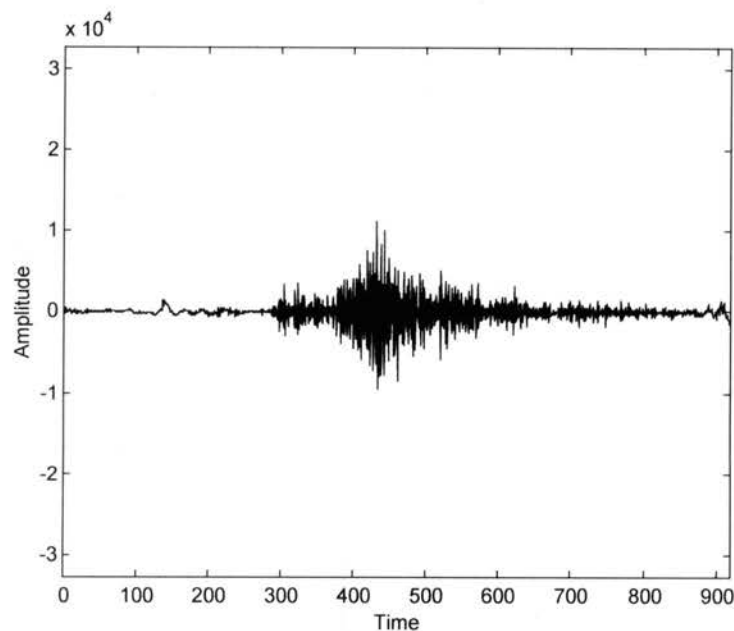


Figure 5-6. All Unvoiced Signal

The third signal shown in Figure 5-7 is a speech signal with no noise and is the word “Figure”. This signal is used to test the response to a segment of speech that is composed of both voiced and unvoiced (mixed) excitation. All three signals are used to test the response of the analysis-by-synthesis loop for the frequency-domain and time-domain approach. The following section describes the sub-sampling process for each of the test signals presented in this section.

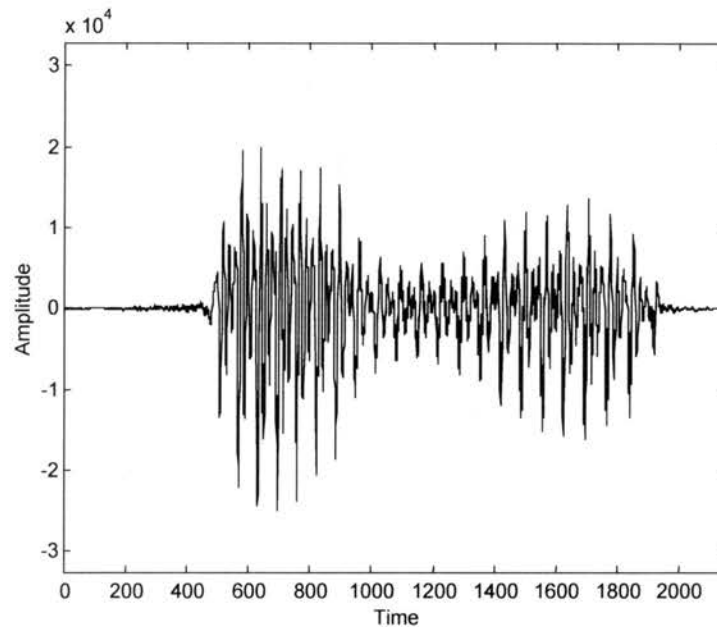


Figure 5-7. The Word “Figure”

5.4.2.2 Sub-Sampling Process

In order to some provide clarity about the sub-sampling process a number of examples are shown below. Figure 5-8 shows the sample points selected in the magnitude spectrum of the all voiced signal when the candidate sub-sampling period is equal to 20 samples. This candidate sub-sampling period produces only 9 sample points, which are designated by the ‘x’. This is clearly not enough sample points to produce an accurate

representation of the original input magnitude spectrum. It is also worth noting that none of the sample points is near the peaks where most of the energy is contained.

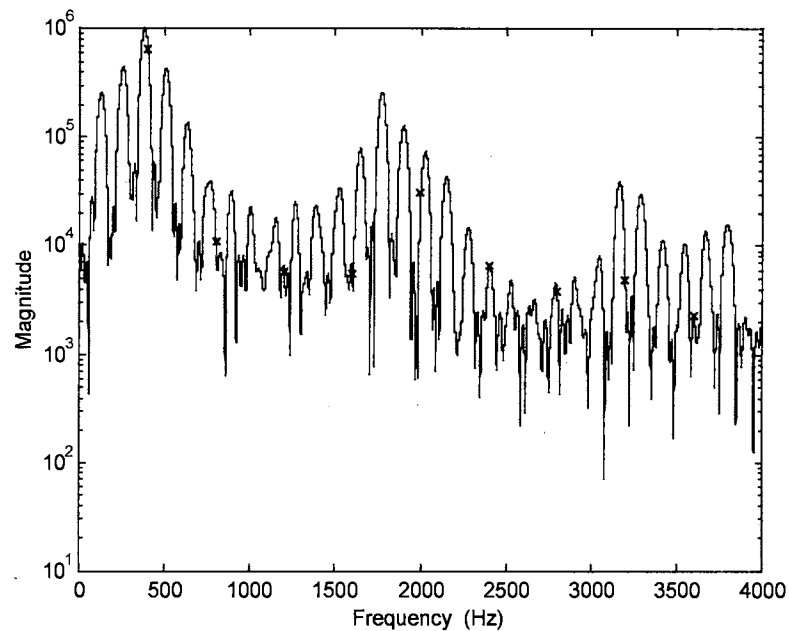


Figure 5-8. All Voiced Magnitude Spectrum Sub-Sampled with $P_s = 20$

The second example of the sub-sampling process is provided in Figure 5-9. The candidate sub-sampling period in this case is 114 samples. The number of sample points that correspond to this sub-sampling period is 56. The larger number of sample points appears to result in a closer approximation to the original magnitude spectrum but the sample points still do not occur near the spectral peaks where the largest energy concentrations are located.

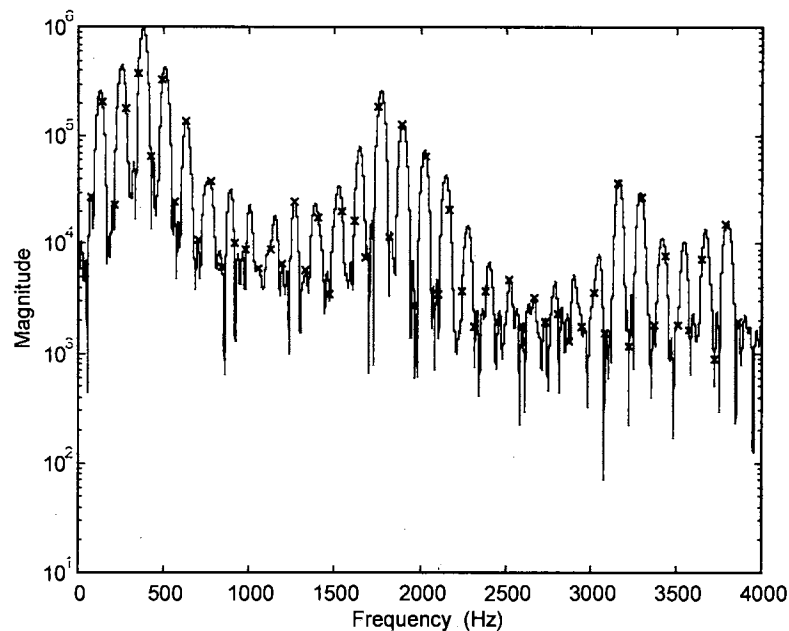


Figure 5-9. All Voiced Magnitude Spectrum Sub-Sampled with $P_s = 114$

The three Figures 5-10, 5-11 and 5-12 below represent how the sample points change as the optimum sub-sampling period is approached and then passed over. Figure 5-10 shows that the sampling points are starting to select the spectral peaks that are present in the magnitude spectrum. But, notice the high frequency region is not being sampled at the appropriate spectral peaks. This error in matching the peaks of the magnitude spectrum is comparable to the error associated with selecting a pitch that is close to the correct pitch but is off by a fraction of a Hertz as is associated with the STC and MBE models. For these models, if the pitch is off even by a fractional value then the error in the sampling process has a multiplicative effect as the magnitude spectrum is being sampled.

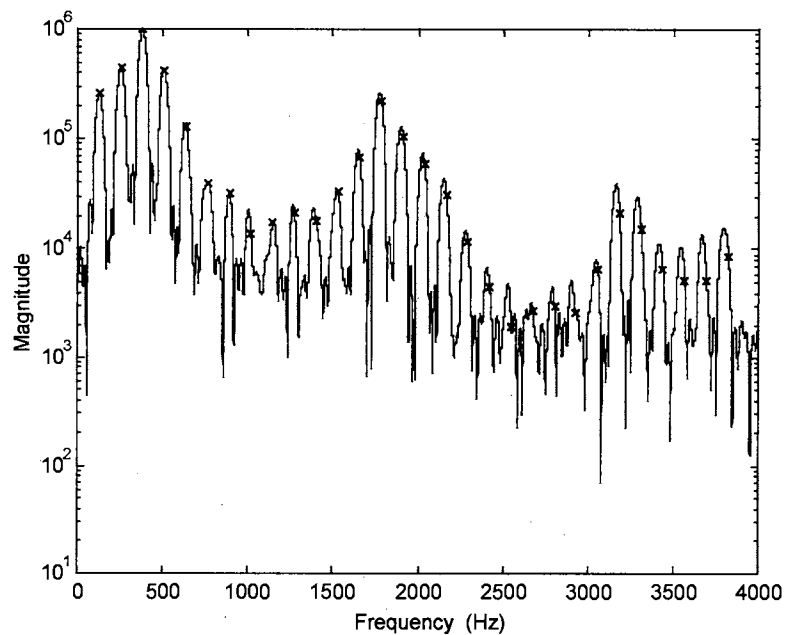


Figure 5-10. All Voiced Magnitude Spectrum Sub-Sampled with $P_s = 62.80$

Figure 5-11 represents the candidate sub-sampling period that generates a magnitude spectrum which corresponds “best” to the magnitude spectrum of the original input speech signal. Notice that each of the sample points select all the spectral peaks in the magnitude spectrum independent of the frequency region.

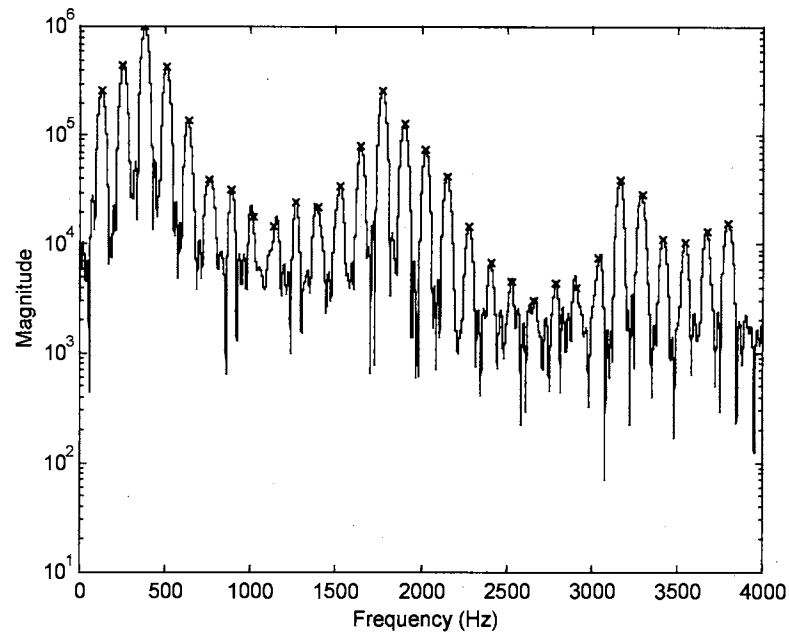


Figure 5-11. All Voiced Magnitude Spectrum Sub-Sampled with $P_s = 63.20$

As the candidate sub-sampling period is incremented past the “best” sub-sampling period the spectral peaks in the high frequency region of the magnitude spectrum are no longer represented accurately. This is seen in Figure 5-12.

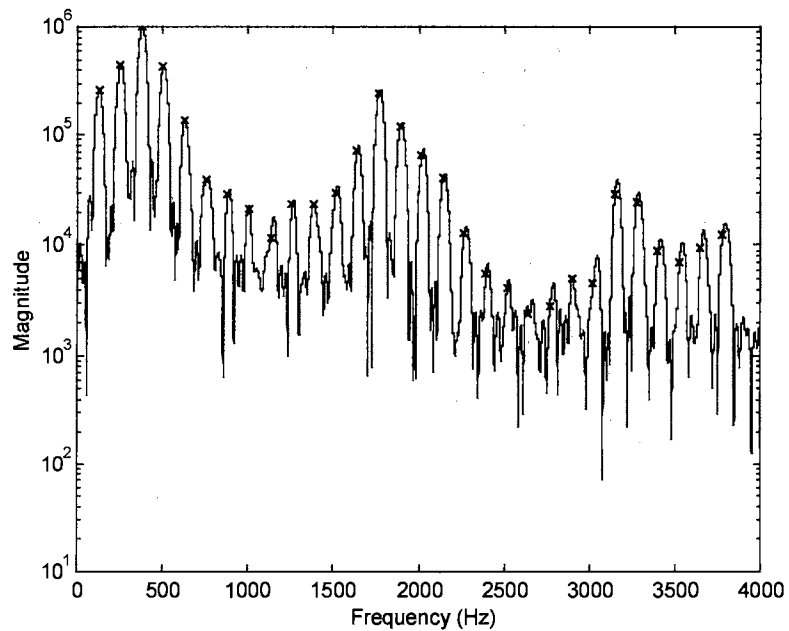


Figure 5-12. All Voiced Magnitude Spectrum Sub-Sampled with $P_s = 63.40$

Figure 5-13 shows the sample points selected in the magnitude spectrum of the all unvoiced signal when the candidate sub-sampling period is equal to 20 samples. Again, this is clearly not enough sample points to produce an accurate representation of the original input magnitude spectrum. The idea is still to model the original magnitude spectrum with minimum error. So even for the unvoiced signal it is important to represent the high energy frequencies in the sub-sampling process.

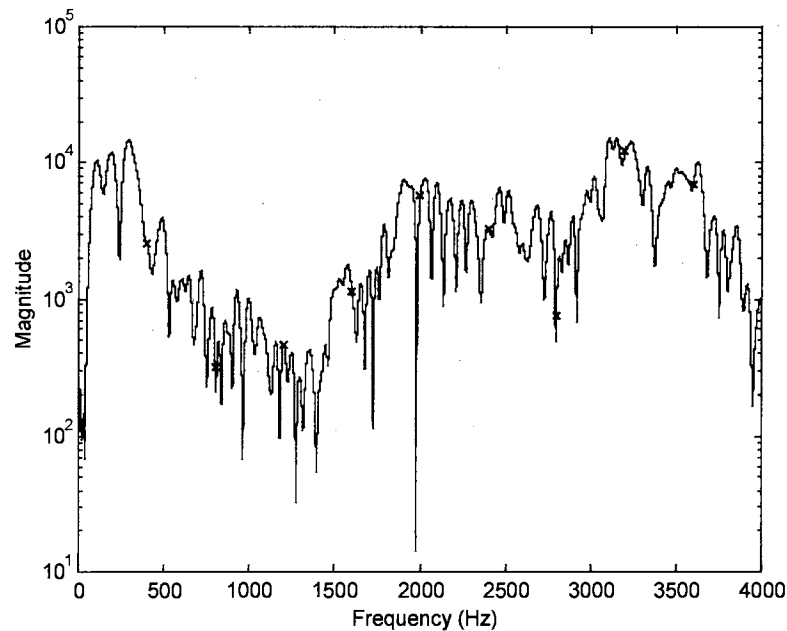


Figure 5-13. All Unvoiced Magnitude Spectrum Sub-Sampled with $P_s = 20$

The second example of the sub-sampling process for the all unvoiced signal is provided in Figure 5-14. The candidate sub-sampling period in this case is 114 samples. Again, the larger number of sample points appears to result in a close approximation to the original magnitude spectrum. In contrast to the all voiced signal the sample points for a sub-sampling period of 114 provides a much closer representation to the original magnitude spectrum. This seems to be appropriate since a large number of sinusoids are necessary to generate a noise signal.

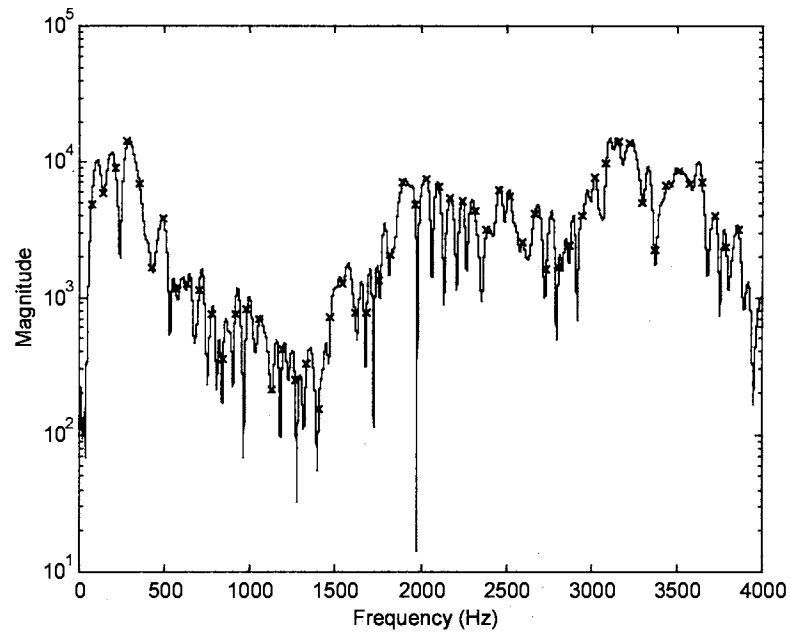


Figure 5-14. All Unvoiced Magnitude Spectrum Sub-Sampled with $P_s = 114$

A third example of the sub-sampling process for the all unvoiced signal is provided in Figure 5-15. The candidate sub-sampling period in this case is 75 samples. Again, the larger number of sample points, compared to the sub-sampling period of 20, appears to result in a close approximation to the original magnitude spectrum.

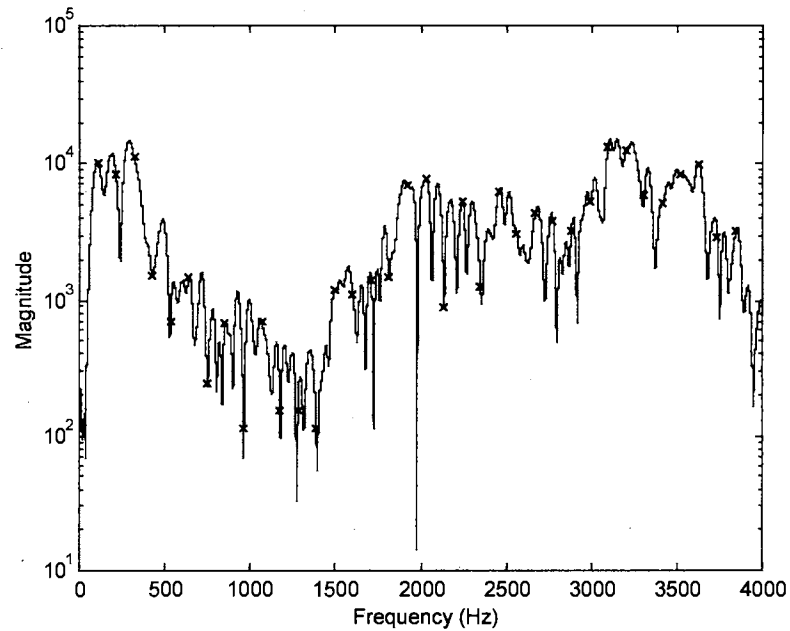


Figure 5-15. All Unvoiced Magnitude Spectrum Sub-Sampled with $P_s = 75$

Figure 5-16 shows the sample points selected in the magnitude spectrum of the mixed excitation signal when the candidate sub-sampling period is equal to 20 samples. Again, this is clearly not enough sample points to produce an accurate representation of the original input magnitude spectrum.

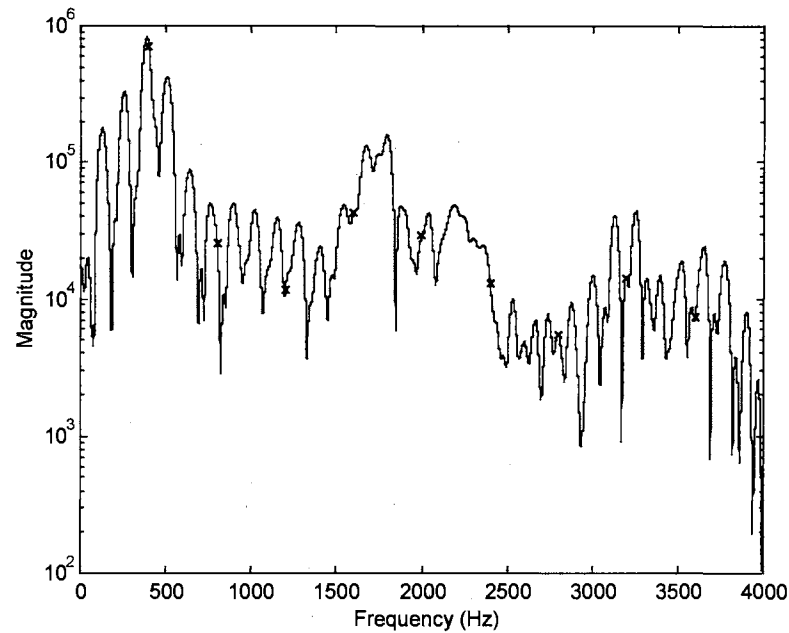


Figure 5-16. Speech Magnitude Spectrum Sub-Sampled with $P_s = 20$

The second example of the sub-sampling process for the mixed excitation signal is provided in Figure 5-17. The candidate sub-sampling period in this case is 114 samples. The large number of sample points does not result in a close approximation to the original magnitude spectrum. In the three test signals, it is obvious that an arbitrarily high sub-sampling period does not result in the appropriate selection of spectral amplitudes.

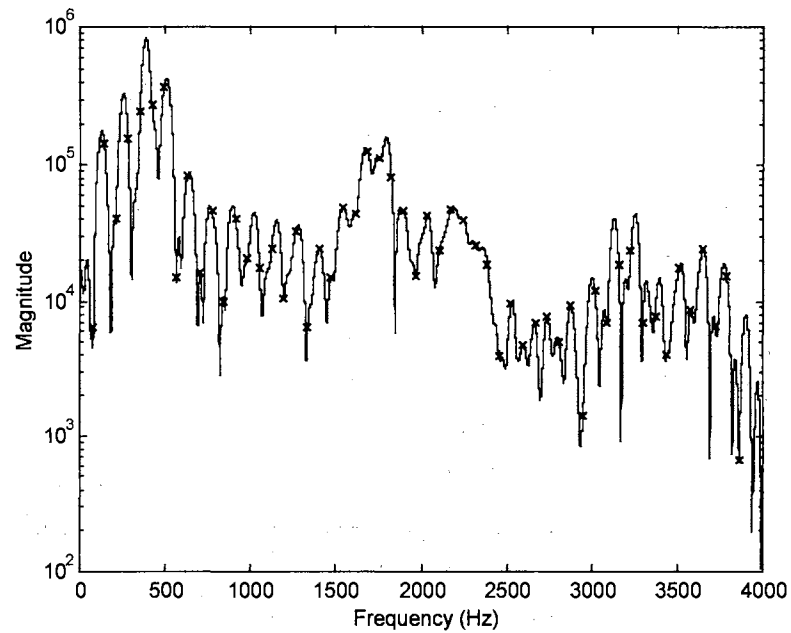


Figure 5-17. Speech Magnitude Spectrum Sub-Sampled with $P_s = 114$

Figures 5-18 and 5-19 below represent how the sample points change as the optimum sub-sampling period is approached for a mixed excitation signal. Figure 5-18 shows that the sampling points are starting to select the spectral peaks that are present in the magnitude spectrum. The low frequency and high frequency regions appear to be sampled properly but the mid-frequency region is not being sampled properly. In Figure 5-19 the low and mid-frequency regions are sampled properly but the high frequency region is not sampled properly.

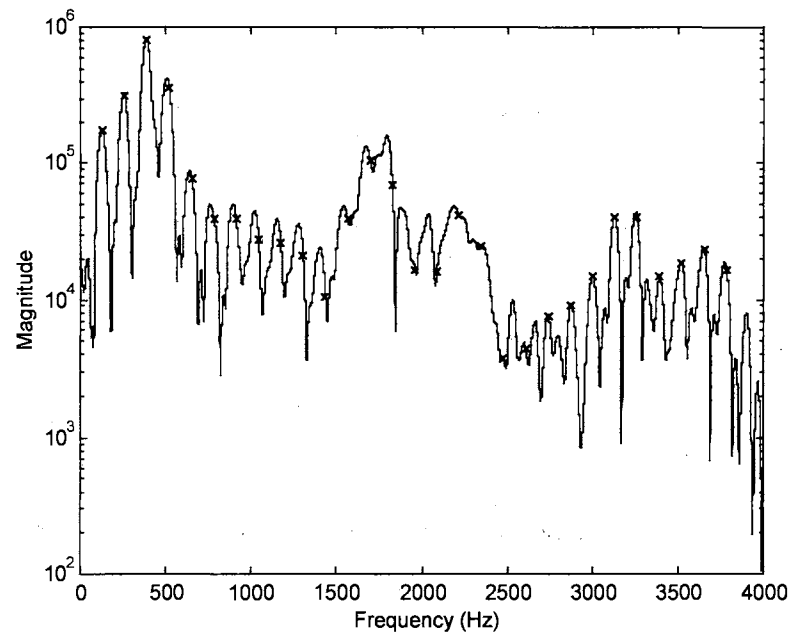


Figure 5-18. Speech Magnitude Spectrum Sub-Sampled with $P_s = 61.2$

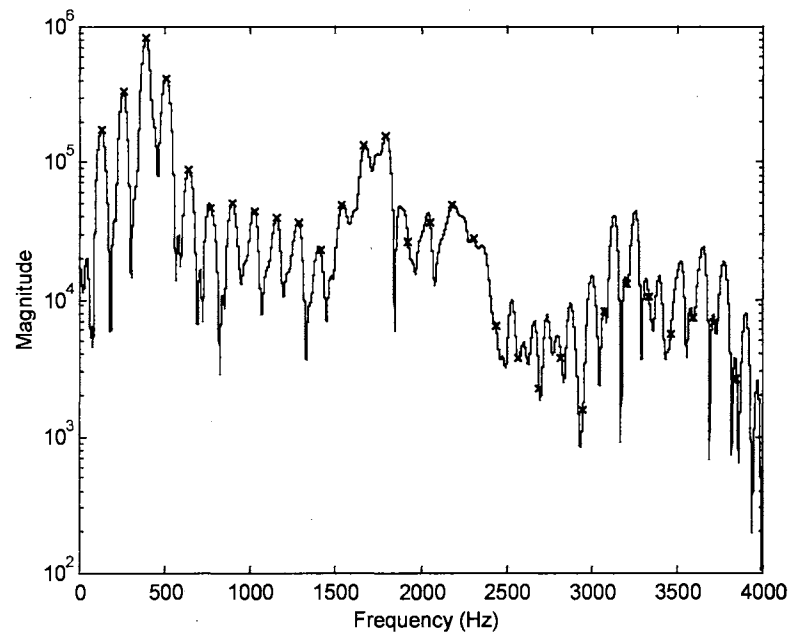


Figure 5-19. Speech Magnitude Spectrum Sub-Sampled with $P_s = 62.4$

5.4.2.3 Analysis-By-Synthesis Loop

This section describes in more detail the operation of the analysis-by-synthesis loop. A comparison of the original magnitude spectrum with the synthetic magnitude spectrum is provided for a particular frame of all three of the test signals.

The first test signal is the all voiced constant tone. Figure 5-20 displays the original magnitude spectrum in the top plot and the synthetic magnitude spectrum for a frame of the all voiced signal at a sub-sampling period of 20 samples. The sub-sampling period of 20 samples and the corresponding spectral amplitudes do not produce an appropriate magnitude spectrum as is clearly obvious. The original magnitude spectrum has approximately 30 spectral peaks and the synthetic magnitude spectrum only has 9 spectral peaks.

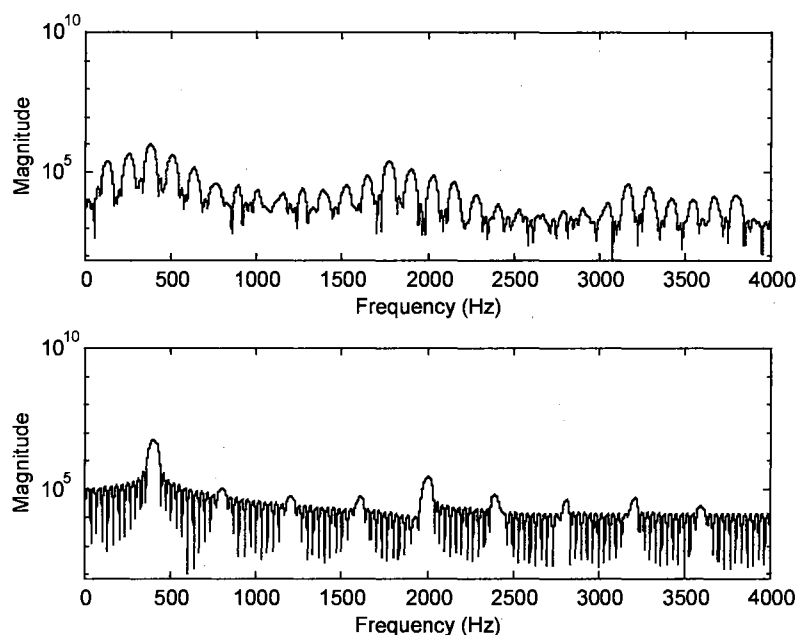


Figure 5-20. Original Magnitude Spectrum and All Voiced Synthetic Magnitude Spectrum for $P_s = 20$

Figure 5-21 shows the original magnitude spectrum in the top plot and the synthetic magnitude spectrum for a frame of the all voiced signal at a sub-sampling

period of 114 samples. The sub-sampling period of 114 samples and the corresponding spectral amplitudes do not produce an appropriate magnitude spectrum although the higher sub-sampling rate produces a closer approximation than for a sub-sampling period $P_s = 20$.

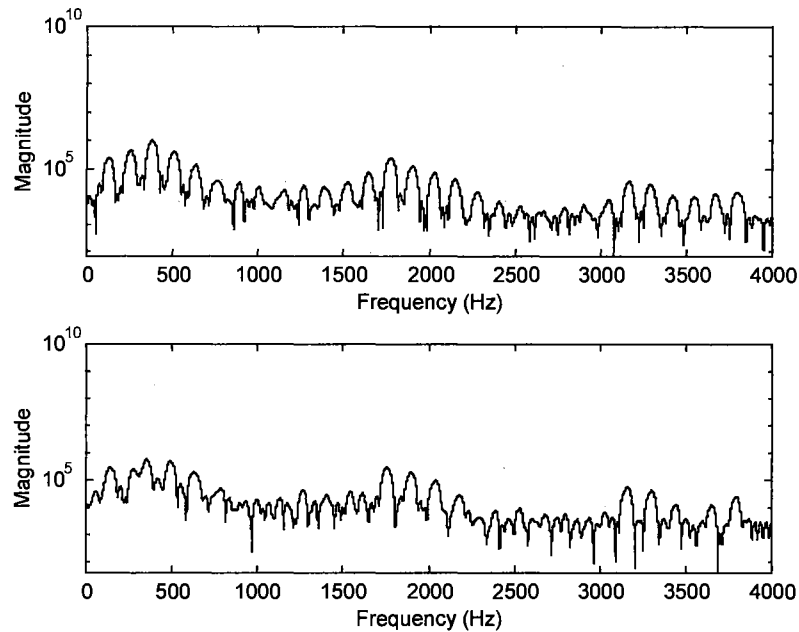


Figure 5-21. Original Magnitude Spectrum and All Voiced Synthetic Magnitude Spectrum for $P_s = 114$

Figures 5-22, 5-23, and 5-24 show the convergence of the analysis-by-synthesis loop. As the optimum sub-sampling period is approached the synthetic all voiced magnitude spectrum converges to the original magnitude spectrum. The synthetic all voiced magnitude spectrum corresponding to $P_s = 62.76$ is shown at the bottom of Figure 5-22. This synthetic magnitude spectrum matches the original magnitude spectrum best in the formant regions but does not match well in the frequency ranges from 800 to 1,200 Hz and from 2,400 to 2,800 Hz. The synthetic all voiced magnitude spectrum for $P_s = 63.2$ is presented at the bottom of Figure 5-23. Again the best matches between the

synthetic magnitude spectrum and the original magnitude spectrum occur in the formant regions. The high frequency region for this sub-sampling period does appear to have better shape over the previous candidate. As the sub-sampling period is incremented to $P_s = 63.4$, the synthetic all voiced magnitude spectrum shown at the bottom of Figure 5-24 again matches best in the formant regions. The high frequency region for this sub-sampling period does not appear to match as well as the previous sub-sampling period. The process of selecting the synthetic all voiced magnitude spectrum that best matches the original is very difficult, which is the reason for the development of the objective measure presented in this dissertation.

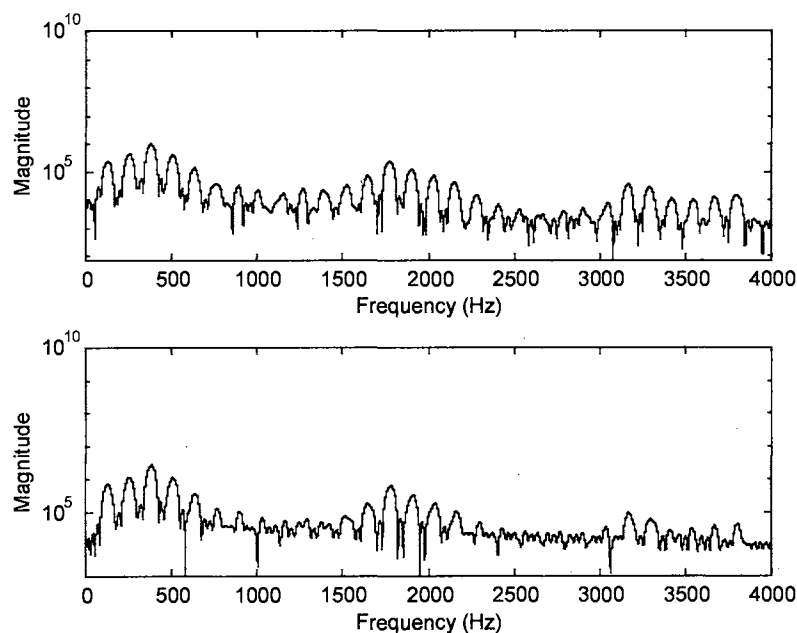


Figure 5-22. Original Magnitude Spectrum and All Voiced Synthetic Magnitude Spectrum for $P_s = 62.80$

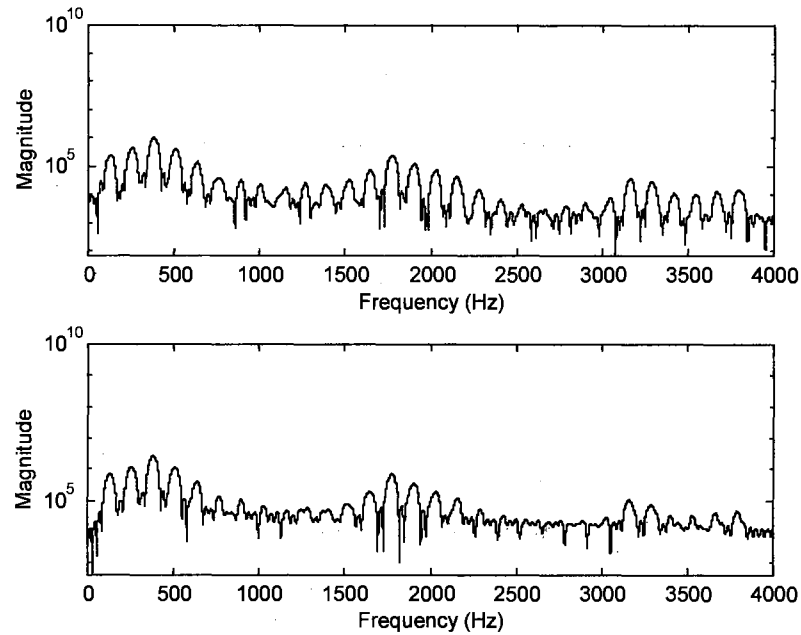


Figure 5-23. Original Magnitude Spectrum and All Voiced Synthetic Magnitude Spectrum for $P_s = 63.2$

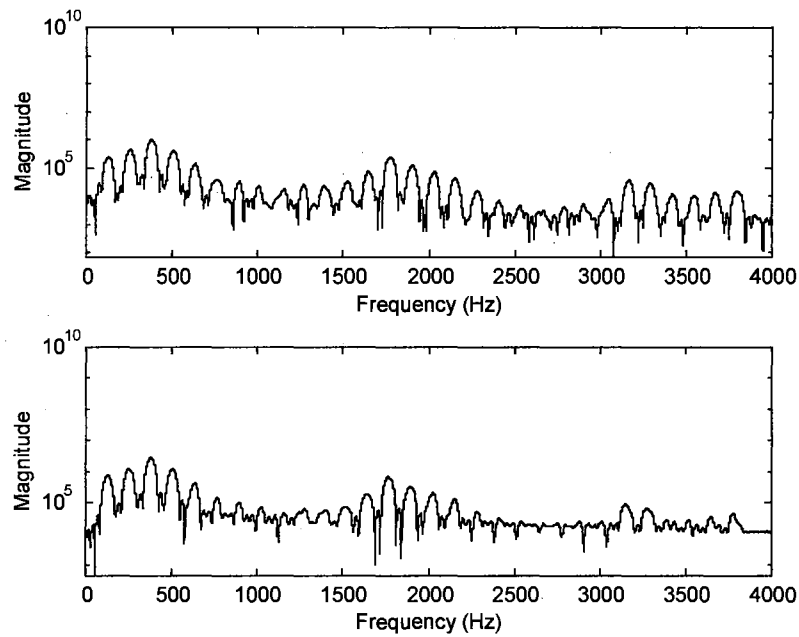


Figure 5-24. Original Magnitude Spectrum and All Voiced Synthetic Magnitude Spectrum for $P_s = 63.4$

The next three figures show the relationship between the original all unvoiced magnitude spectrum and the synthetic all unvoiced magnitude spectrum for sub-sampling

periods of $P_s = 20$, $P_s = 114$, and $P_s = 75$, shown in Figures 5-25, 5-26, and 5-27, respectively. The synthetic magnitude spectra in Figures 5-25 and 5-27 do not represent accurately the original magnitude spectrum while the synthetic magnitude spectrum of Figure 5-26 appears to be a good match. The synthetic all unvoiced magnitude spectrum for $P_s = 20$ looks like a voiced spectrum because of the low number of sample points which suggest that the sub-sampling period should be higher. The synthetic all unvoiced magnitude spectrum for $P_s = 75$ again appears to have too much harmonic structure to be a good match for the original magnitude spectrum. As the sub-sampling period is increased, the synthetic all unvoiced magnitude spectrum takes on the shape of the original all unvoiced magnitude spectrum.

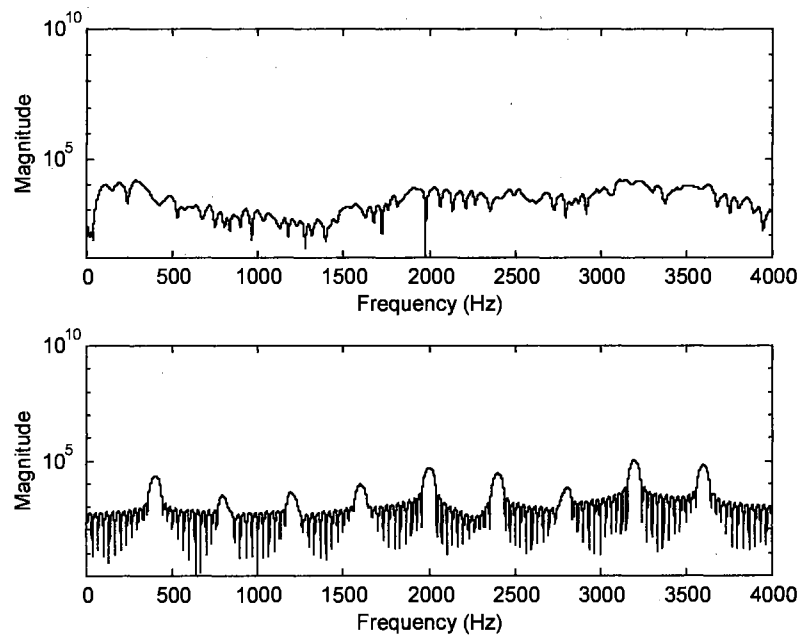


Figure 5-25. Original Magnitude Spectrum and All Unvoiced Synthetic Magnitude Spectrum for $P_s = 20$

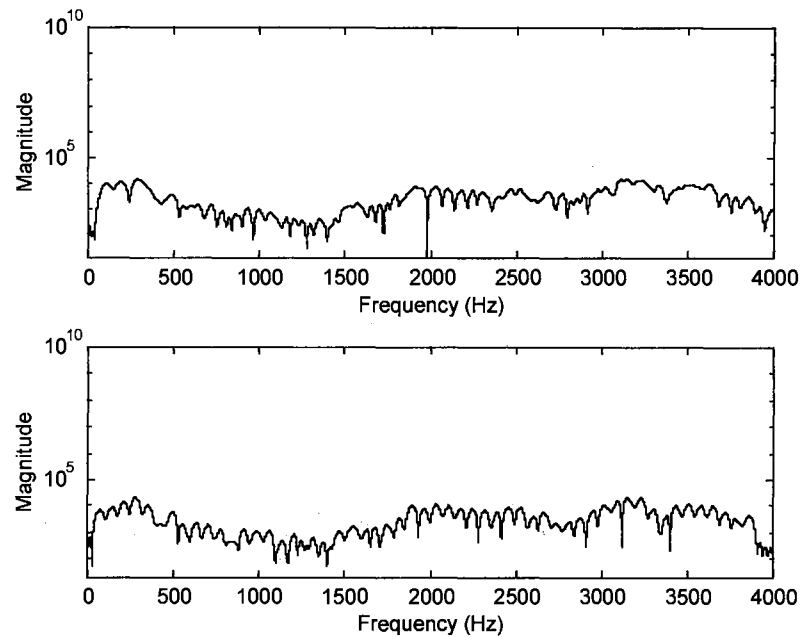


Figure 5-26. Original Magnitude Spectrum and All Unvoiced Synthetic Magnitude Spectrum for $P_s = 114$

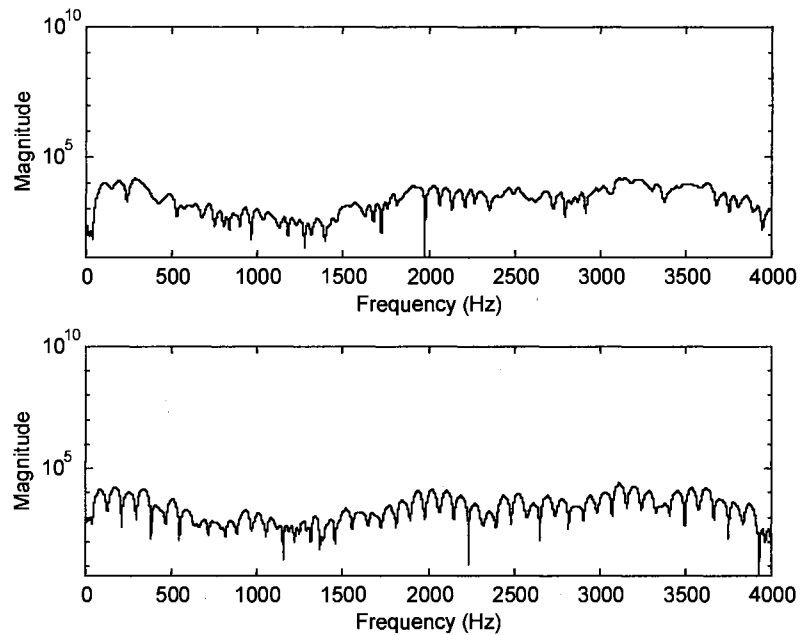


Figure 5-27. Original Magnitude Spectrum and All Unvoiced Synthetic Magnitude Spectrum for $P_s = 75$

The next set of figures is an analysis of the response of the frequency-domain analysis-by-synthesis to a frame of a real speech signal. Once again the synthetic

magnitude spectrum shown in Figure 5-28 and corresponding to $P_s = 20$ does not match the original magnitude spectrum because of the low number of sampling points.

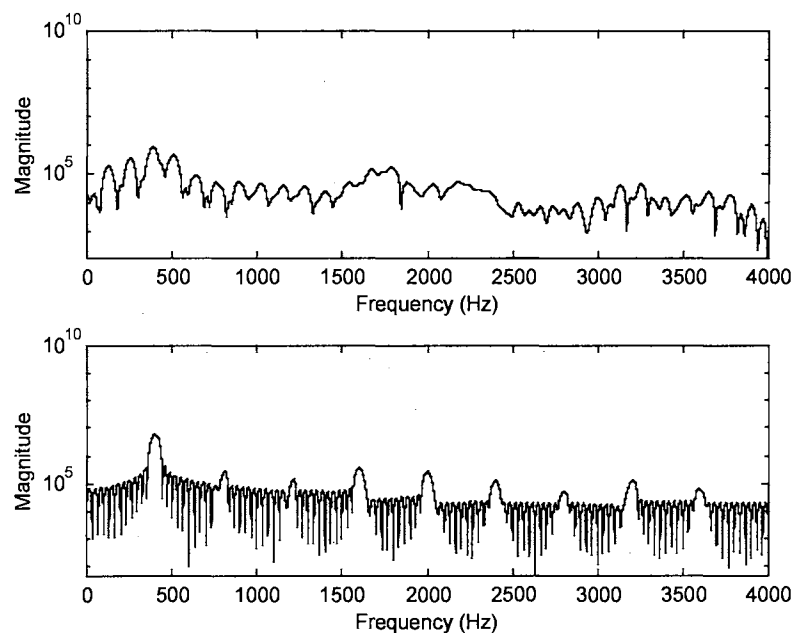


Figure 5-28. Original Magnitude Spectrum and “Figure” Synthetic Magnitude Spectrum for $P_s = 20$

If a synthetic magnitude spectrum is generated for a sub-sampling period of $P_s = 114$, then the magnitude spectrum starts looking like an unvoiced spectrum as shown in Figure 5-29. This is a result of the high number of sample points (56).

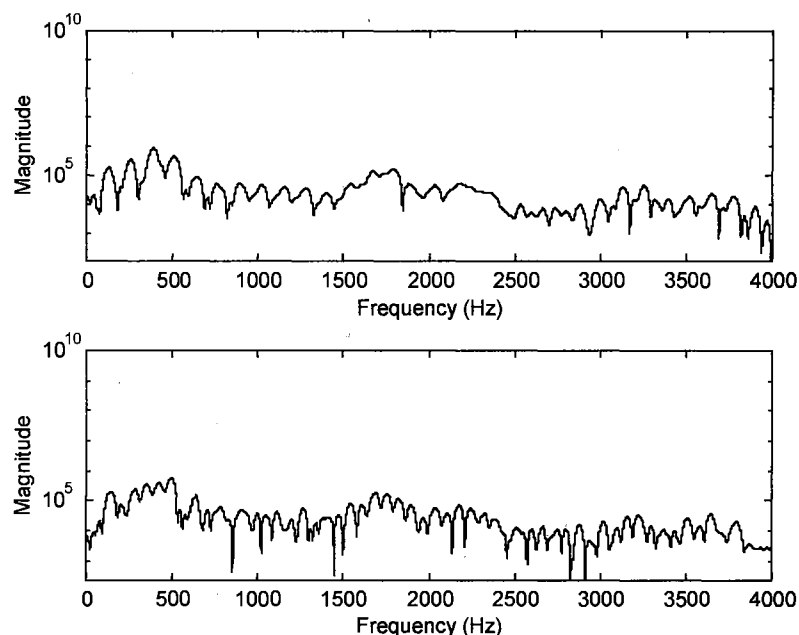


Figure 5-29. Original Magnitude Spectrum and “Figure” Synthetic Magnitude Spectrum for $P_s = 114$

As previously stated, subjectively selecting the synthetic magnitude spectrum that best matches the original magnitude spectrum is a difficult task. This is clearly obvious in the analysis for this particular frame of speech. The original magnitude spectrum appears to have a strong voiced region, an unvoiced region, and a voiced region. The synthetic magnitude spectra corresponding to $P_s = 61.2$ and $P_s = 62.4$ shown in Figures 5-30 and 5-31 exhibit the properties of a frame of speech that is declared entirely voiced. This issue is addressed in the conclusion. Disregarding the potential error in the voicing decisions, both synthetic magnitude spectra are representative of the original magnitude spectrum. The synthetic magnitude spectrum of Figure 5-31 does appear to model the last peak of the original magnitude spectrum better than the synthetic magnitude spectrum of Figure 5-30. The next section discusses the match scores for all three test signals.

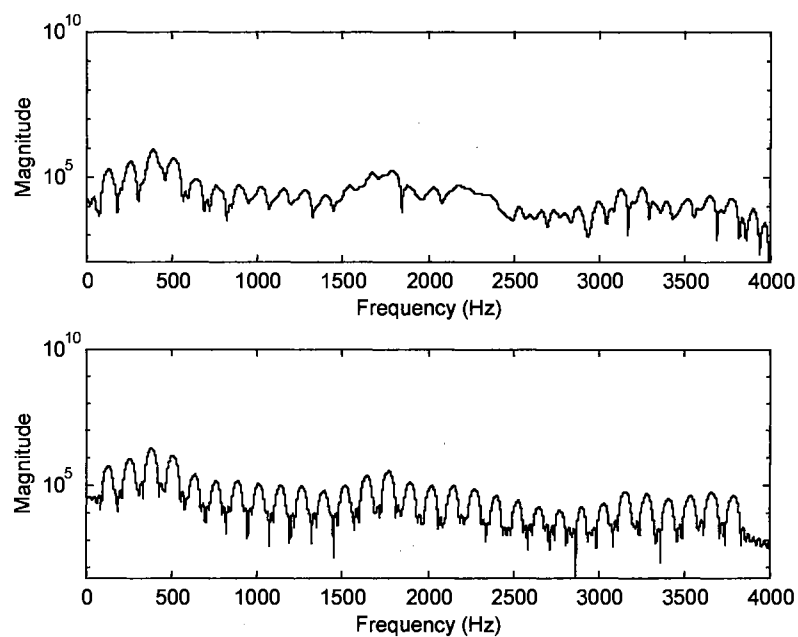


Figure 5-30. Original Magnitude Spectrum and “Figure” Synthetic Magnitude Spectrum for $P_s = 61.20$

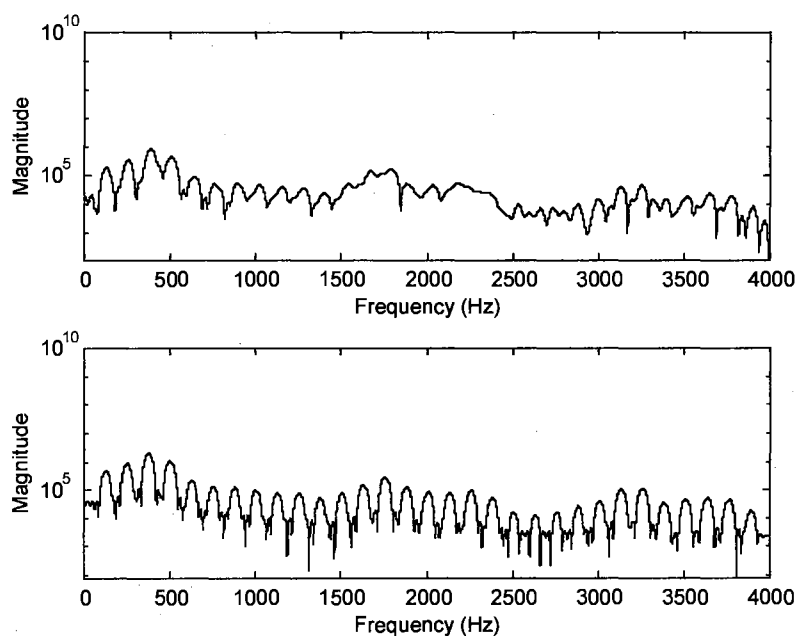


Figure 5-31. Original Magnitude Spectrum and “Figure” Synthetic Magnitude Spectrum for $P_s = 62.40$

5.4.2.4 Match Scores

This section discusses the match scores for the frames of speech that are being analyzed thus far in this chapter. The match scores represent how well the synthetic magnitude spectrum, given an optimum gain, for each candidate sub-sampling period corresponds to the original magnitude spectrum.

The process of finding the optimum sub-sampling period is split into a two stage process because of the sheer computational complexity of the problem (correlation of sequences of length 8,192). The first stage computes the match scores for all the possible integer sub-sampling periods, and stage two performs a refinement around the sub-sampling period producing the highest match score in the first stage. This refinement ranges plus and minus two samples in 0.2 increments. The results for a particular frame are presented below.

The integer match scores for the synthetic all voiced magnitude spectrum is shown in Figure 5-32. Note, there are a number of methods for finding the minimum solution to the mean-squared error problem as developed in section 5.4, but for the methods to work correctly only one minimum should exist. In the case of analyzing speech or speech like signals using analysis-by-synthesis it is clearly seen that these signals have more than one possible minimum value as shown in Figure 5-32 and in the following figures.

The integer match scores for the synthetic all voiced magnitude spectra have 6 possible maxima for this particular frame of the all voiced signal. This is the first indication that the frequency-domain analysis-by-synthesis approach to selecting the appropriate sub-sampling period is correct.

The all voiced signal has a fundamental frequency equal to approximately 126 Hz or 63.4 samples. The sub-sampling period producing the highest match score is 63 samples. It is also worth pointing out that a potential maximum occurs at the integer sub-sampling period of 32 samples, this corresponds to the effect of pitch doubling present in the common open-loop pitch estimation algorithms. The other potential maximum match scores are a result of other sub-sampling periods providing reasonable fits to the original magnitude spectrum in specific frequency ranges.

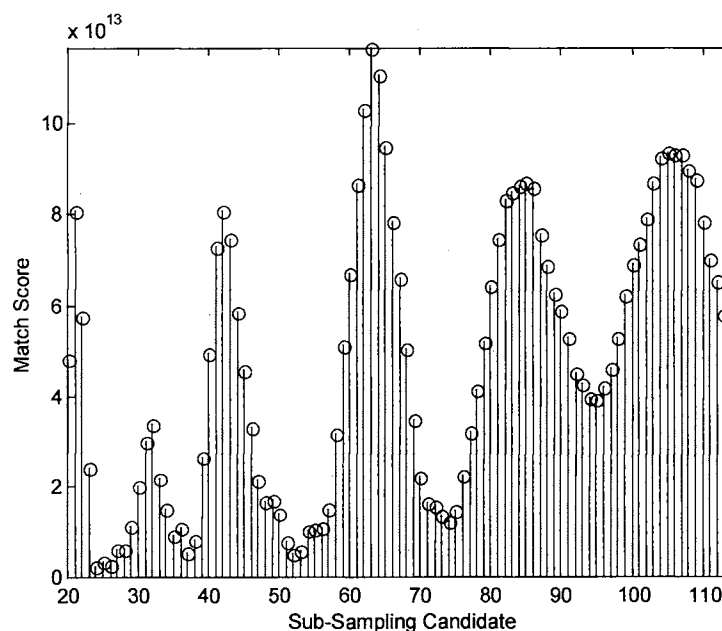


Figure 5-32. Match Scores for Integer Sub-Sampling Periods of the All Voiced Signal

The set of match scores corresponding to the fractional sub-sampling periods is shown in Figure 5-33. The sub-sampling period resulting in the highest match score is 63.2 samples, although the sub-sampling periods of 63 and 63.4 also produce high match scores. This is not exactly equal to the fundamental frequency of the original all voiced signal. The error is associated with the resolution in the DFT. It is commonly known that the longer the DFT the more accurate the frequency resolution of the signal being

analyzed. Typically, the more accurate frequency resolution is obtained by padding the input signal with zeros. In this case, if an even longer DFT is used then the sub-sampling period corresponding to the exact fundamental frequency is probable.

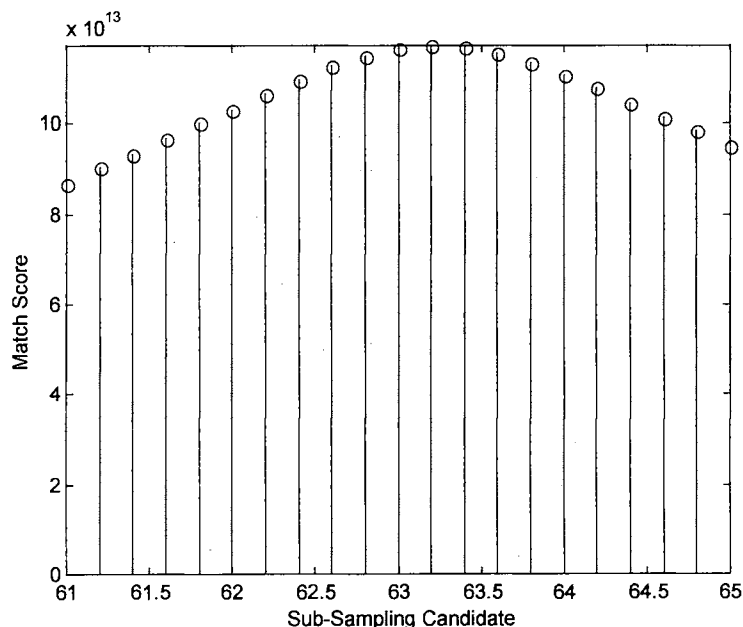


Figure 5-33. Match Scores for Fractional Sub-Sampling Periods of the All Voiced Signal

The match scores corresponding to the synthetic all unvoiced magnitude spectrum are presented in Figures 5-34 and 5-35. As stated previously, the synthetic all unvoiced magnitude spectrum becomes a closer match as the sub-sampling period is increased. This is shown to be the case in Figure 5-34. Notice that as the sub-sampling period is increased there is a steady although not continuous climb in match score. Again, note the number of potential maximum values possible when trying to minimize along the line.

The match scores for the fractional sub-sampling periods are shown in Figure 5-35. In this case the sub-sampling period producing the highest match score is the same as in the integer case. Sub-sampling periods that do not fall within the range of 20 to 114 are not considered to be reliable candidates and are not considered.

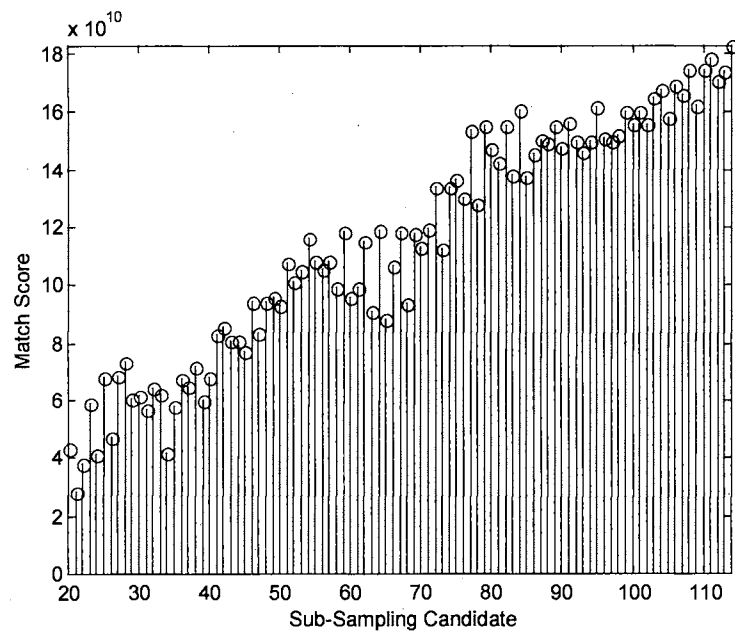


Figure 5-34. Match Scores for Integer Sub-Sampling Periods of the All Unvoiced Signal

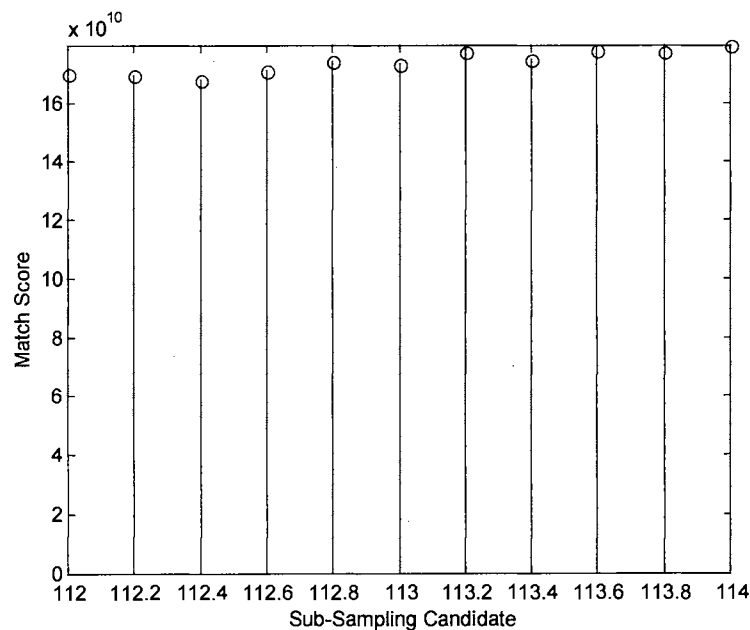


Figure 5-35. Match Scores for Fractional Sub-Sampling Periods of the All Unvoiced Signal

The first two signals, the all voiced and all unvoiced, produce match score contours that produce the appropriate selection of sub-sampling period as is expected for nearly ideal conditions. The final test is to determine the match scores in response to a

frame of a real speech signal, the word “Figure”. The integer sub-sampling period match scores are presented in Figure 5-36 and the fractional sub-sampling match scores are presented in Figure 5-37. Again, there are a number of potential maxima on the match score contour (a line). The integer sub-sampling period producing the highest match score is 62 samples. This sub-sampling period is then refined over the range 60 to 64 samples with $P_s = 62.4$ producing the highest match score. Based on observation and experience with this particular word the sub-sampling period found using the frequency-domain analysis-by-synthesis is a reasonable estimate. The next section discusses the resulting sub-sampling period contour produced after analyzing each of the test signals.

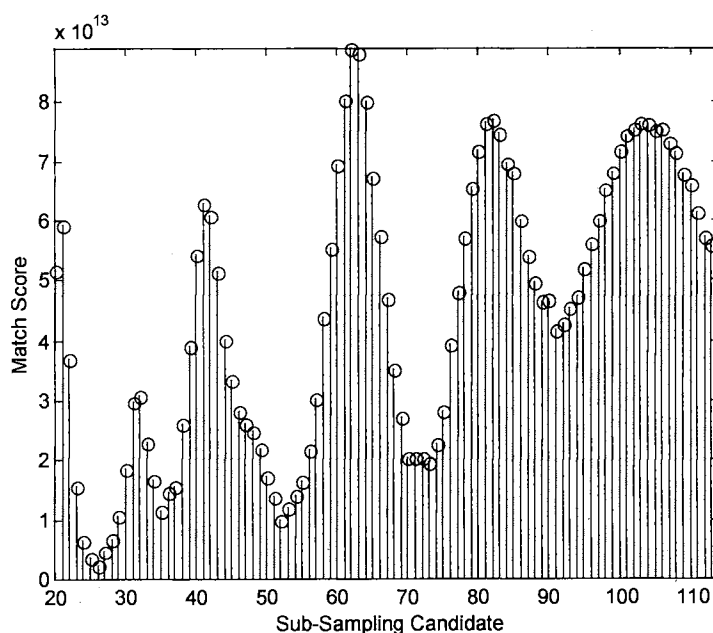


Figure 5-36. Match Scores for Integer Sub-Sampling Periods of the Word “Figure”

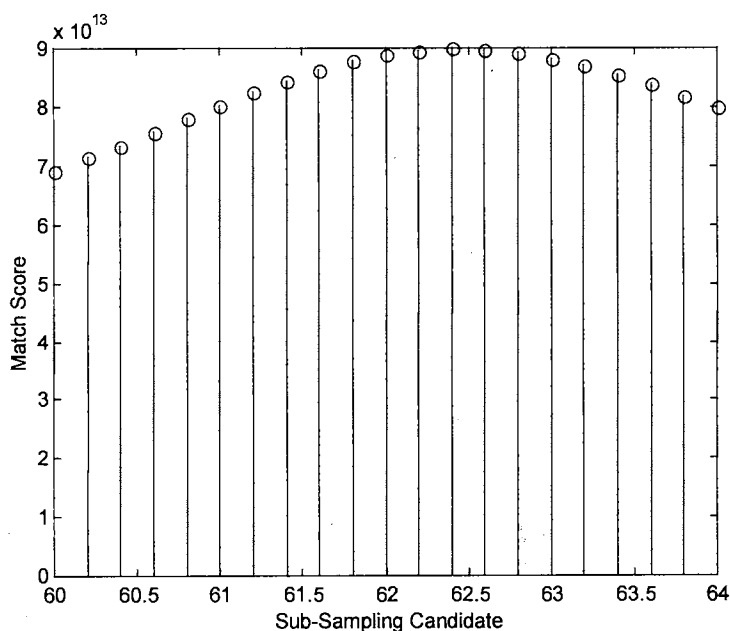


Figure 5-37. Match Scores for Fractional Sub-Sampling Periods of the Word “Figure”

5.4.2.5 Sub-Sampling Period Contour

This section discusses the results of the sub-sampling period contour obtained by analyzing the entire test signals: all voiced, all unvoiced, and the word “Figure”. Since the all voiced signal is a constant tone it is expected that the sub-sampling period contour would also be constant. Figure 5-38 shows the sub-sampling period contour produced using the frequency-domain analysis-by-synthesis method to estimate the sub-sampling period. The contour, as expected, is constant except in the transition regions at the beginning and end of the signal.

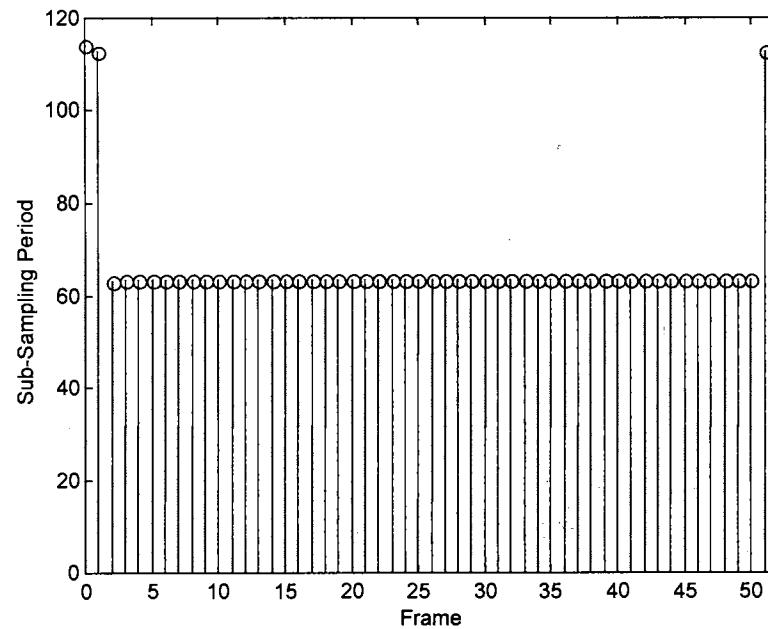


Figure 5-38. All Voiced Sub-Sampling Period Contour

The second test signal is an all unvoiced signal. In contrast to the constant tone it is expected that the sub-sampling period contour would not be constant but vary from frame-to-frame. Also it is expected that the sub-sampling periods would be biased towards the higher sub-sampling periods. Figure 5-39 shows the sub-sampling period contour produced using frequency-domain analysis-by-synthesis to estimate the sub-sampling period. The contour, as expected, is not constant from frame-to-frame and is biased towards the higher sub-sampling periods.

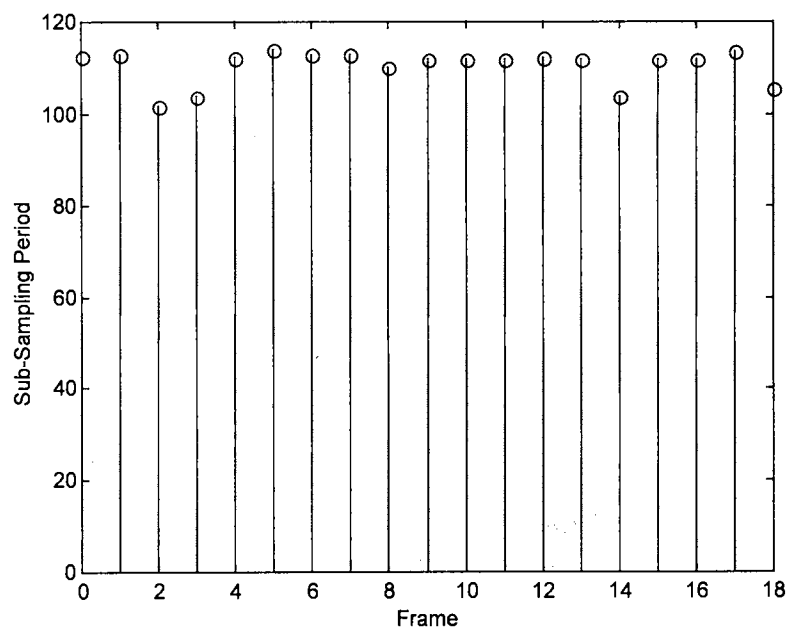


Figure 5-39. All Unvoiced Sub-Sampling Period Contour

The third test signal and the real key to the success of the frequency-domain analysis-by-synthesis method is a real speech signal. Unlike the all voiced signal, which is a constant tone, the speech signal in general is a time-varying signal so the fundamental frequency varies from frame-to-frame. The sub-sampling period contour is expected to vary slightly from frame-to-frame in the areas of voiced speech and be biased towards the higher sub-sampling periods. Figure 5-40 shows the sub-sampling period contour produced using frequency-domain analysis-by-synthesis to estimate the sub-sampling period. The contour, as expected, is not constant from frame-to-frame but tracks the time-varying properties of the speech signal from frame-to-frame in the voiced regions. In the transition and unvoiced regions the sub-sampling period contour is biased towards the higher sub-sampling periods. The next section describes the synthetic signals produced from parameter estimates obtained by the frequency-domain analysis-by-synthesis method.

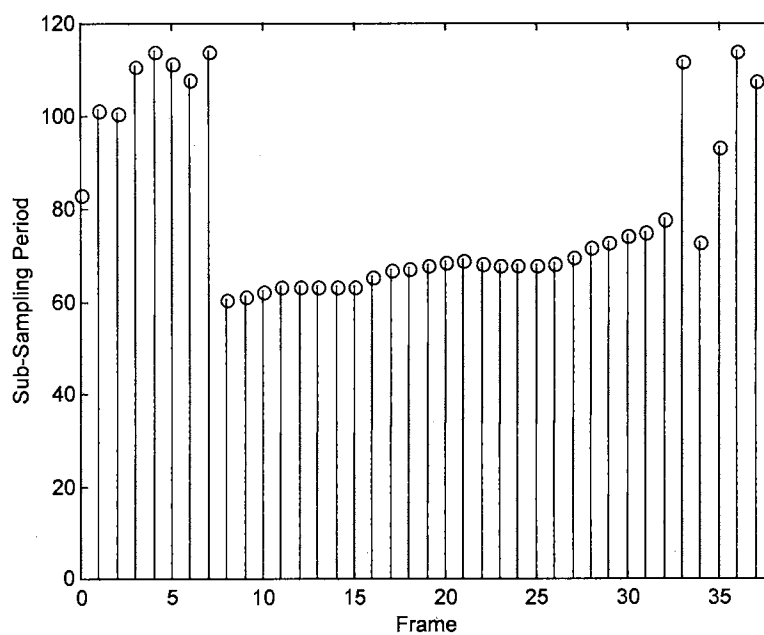


Figure 5-40. “Figure” Sub-Sampling Period Contour

5.4.2.6 Synthesized Test Signals

This section looks at the resulting synthetic signals produced from the parameter estimates obtained from the frequency-domain analysis-by-synthesis approach and using the sinusoidal model for reconstructing the synthetic signal.

The three test signals, all voiced, all unvoiced, and the word “Figure”, are shown in Figures 5-41, 5-42, and 5-43, respectively. The synthetic all voiced signal is a close approximation to the original all voiced signal shown in Figure 5-5. The main differences are that the synthetic signal is delayed, the onset is not as sharp, and the amplitude modulation is slightly enhanced. There is a difference in the maximum and minimum values in the synthetic signal but this is dismissed because the reconstructed signal is guaranteed to be out of phase with the original signal thus producing the difference in maximum and minimum values.

The synthetic all unvoiced signal is not a good approximation to the original all unvoiced signal. Again the reconstructed signal is delayed and the maximum and minimum values are different. One interesting note is that the reconstructed signal appears to have more structure than the original all unvoiced signal. This is attributed to the phase difference between the original and the reconstructed.

The reconstructed signal for the word “Figure” does not resemble the original signal. The synthetic signal is delayed, the noise regions are attenuated, and the harmonic structure is stronger in the voiced regions. The latter two are attributed to the fact that the phase relationship between the synthetic and original signal is different. While all three of the synthetic test signals have a few differences, the true test is the listening test. If the synthetic speech looks almost exactly like the original but does not sound anything like the original then the method would not be useful for analysis or synthesis. If the synthetic speech is similar to the original and sounds like the original then the method is useful for analysis and synthesis. In all three cases the synthetic signal sounds similar to the original signal.

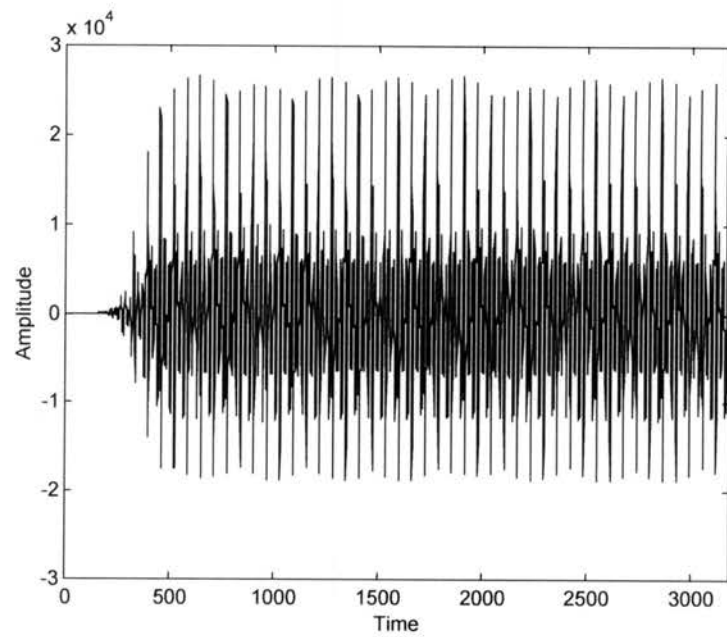


Figure 5-41. Synthetic All Voiced Signal

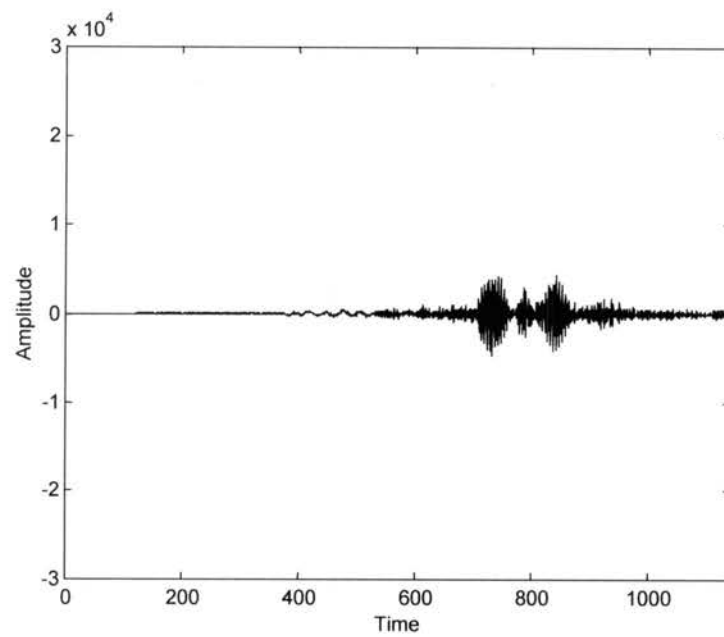


Figure 5-42. Synthetic All Unvoiced Signal

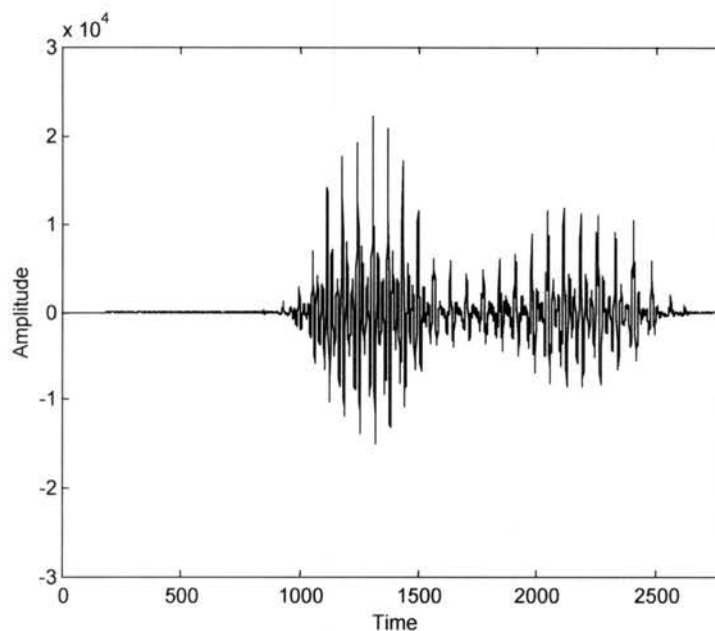


Figure 5-43. Synthetic Word “Figure”

5.4.2.7 Conclusion

In section 5.4 a frequency-domain analysis-by-synthesis method of selecting the appropriate parameters for the sinusoidal model was developed. A complete theoretical development was presented along with the simulation results.

The simulation was tested on three different test signals: an all voiced, an all unvoiced, and a speech signal (the word “Figure”). The all voiced signal is used to test the response of the system on a constant tone. In contrast to the all voiced signal, the all unvoiced signal is used to test the system response to a noise signal. Both of these represent the ideal conditions for pure voiced speech and pure unvoiced speech. The word “Figure” is added to test the system response to a more realistic signal.

The frequency-domain analysis-by-synthesis method is shown to respond as expected given the signals tested. A major disadvantage to the frequency-domain approach is the amount of computational complexity required to perform the analysis-by-

synthesis loop. The DFT length chosen is 16,384 points. The analysis-by-synthesis loop has to perform correlations on sequences of length 8,192. This is done for 95 integer sub-sampling periods and a maximum of 21 fractional sub-sampling periods. A second weakness is in tying the voicing decisions into the analysis-by-synthesis loop, as is evident in the reconstruction of the synthetic all voiced magnitude spectrum shown in Figures 5-22, 5-23, and 5-24.

The next section introduces an alternate approach to the frequency-domain analysis-by-synthesis approach presented in this section. The alternate approach is a time-domain analysis-by-synthesis method.

5.5 Time-Domain Analysis-By-Synthesis Sinusoidal Model

5.5.0 Introduction

This section describes the development of the time-domain analysis-by-synthesis method. The analysis and synthesis techniques described in sections 5.2 and 5.3 are combined in a closed-loop fashion to estimate the model parameters using a mean-squared error approach similar to that described in the frequency-domain section. The time-domain approach differs from the frequency-domain approach in that the assumption is that phase information is transmitted, thus the phase information is available in the analyzer and the synthesizer. A block diagram for the time-domain analysis-by-synthesis approach is shown in Figure 5-5.

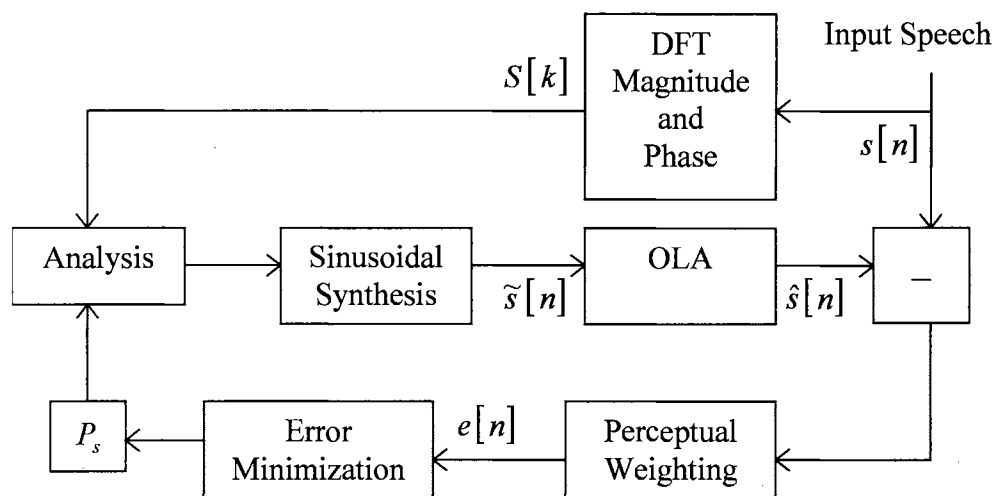


Figure 5-44. Time-Domain Analysis-By-Synthesis Using Sinusoidal Model

First, the input vector is analyzed by computing the magnitude and phase response of the DFT. Then a candidate sub-sampling period P_s is selected to represent the current analysis frame. Using the technique described in section 5.2 the amplitudes and phases are estimated as a function of the candidate sub-sampling period. The parameters needed to synthesize speech using the sinusoidal model are sub-sampling period, phase, and spectral amplitudes. Using these parameter estimates, the sinusoidal model is used to generate a frame of synthetic speech. The synthetic speech vector is subtracted from the input speech vector and a MSE value is computed. The MSE is used to select another sampling period, which selects an alternate set of amplitudes and phases. This process is repeated until a MSE value is computed for each candidate sub-sampling period. The sub-sampling period that selects the set of model parameters having the minimum MSE is chosen to represent the current frame of input speech.

The search procedure consists of finding a sub-sampling period that produces the set of parameters that in turn produces synthetic speech that best matches the input speech in a weighted least square error sense. The next section describes the

mathematical development for the time-domain analysis-by-synthesis procedure used to determine the parameters for the sinusoidal model of reconstruction.

5.5.1 Time-Domain Analysis-By-Synthesis

The time-domain analysis-by-synthesis method developed in this section is nearly equivalent to the frequency-domain analysis-by-synthesis method derived in section 5.4; however the two approaches are completely different. One method is derived in the frequency-domain using no phase information and the other method is derived in the time-domain using phase information. There are some new assumptions made about the synthesis model, which are pointed out during the development.

As in the frequency-domain approach, the sinusoidal model is used to generate a frame of synthetic speech. The synthetic speech vector is subtracted from the input speech vector and a mean-squared error value is computed. The MSE is used to select another candidate sub-sampling period, which selects an alternate set of spectral amplitudes and phases. This process is repeated until a MSE value is computed for each candidate sub-sampling period. The sub-sampling period that corresponds to the set of model parameters having the minimum MSE are chosen to represent the current frame of input speech.

Again, the MSE procedure is written mathematically as

$$E = \frac{1}{N} \sum_{n=0}^{N-1} e^2[n] \quad (5-46)$$

where $e[n] = s_{w_r}[n] - \hat{s}[n]$, $s_{w_r}[n]$ is the appropriately windowed input signal, and $\hat{s}[n]$ is given by equation 5-16. The input signal has the reconstruction window $w_r[n]$ applied so

that the comparison between the input signal $s[n]$ and the synthetic signal $\hat{s}[n]$ is approximately a one-to-one matching. The total error E after substituting equation 5-16 into equation 5-46 and substituting in the reconstruction window, $w_r[n]$, is written as

$$E = \frac{1}{N} \sum_{n=0}^{N-1} \left(s[n]w_r[n] - \left(\tilde{s}^{k-1}[n]w_r[n] + \tilde{s}^k[n]w_r[n-N] \right) \right)^2. \quad (5-47)$$

In the frequency-domain approach equation 5-47 is not used because of the lack of phase information which led to a time alignment between the original input signal and the synthetic signal. For the time-domain approach equation 5-47 is valid since the phase information is being used in the synthesizer.

As in the frequency-domain approach, this problem is viewed as an extremely complex problem to solve because of the number of time-varying parameters and the dependence on knowing information about frequency-domain parameters using a time-domain synthesis. This problem is simplified by redefining equation 5-16, making a number of assumptions about the generation of the synthetic signal, and performing the analysis-by-synthesis in the time-domain. Noting that the energy in the synthetic signal is not equal to the energy in the original signal as a result of the sub-sampling process, a gain term is introduced into equation 5-20. Equation 5-20 is presented again here for clarity as given by

$$\hat{s}[n] = w_r[n]g^{k-1}\tilde{s}^{k-1}[n] + w_r[n-N]g^k\tilde{s}^k[n-N]. \quad (5-48)$$

Equation 5-47 is now rewritten by substituting equation 5-48 producing the new total error E as defined by

$$E = \frac{1}{N} \sum_{n=0}^{N-1} \left(s[n]w_r[n] - \left(g^{k-1}\tilde{s}^{k-1}[n]w_r[n] + g^k\tilde{s}^k[n]w_r[n-N] \right) \right)^2. \quad (5-49)$$

This is exactly the same problem that is being solved in the frequency-domain analysis-by-synthesis approach, only now it is solved in the time-domain. This equation still does not lend itself to an easy minimum solution in either the time-domain or frequency-domain. But now the time-domain approach no longer has problems with misalignment because of the inclusion of phase information. This is the point where the two approaches start to diverge at least for the moment.

The total error E to minimize is as defined in equation 5-49. The frequency-domain approach had to be simplified to not include the overlap portion of the reconstruction. The time-domain approach is going to use the overlap from the previous frame to help maintain the most accurate representation of the time-domain signal as possible.

Equation 5-49 is now rewritten in a form that is more usable. The signal and gain from the previous frame are considered to form a known sequence and is represented by $\hat{s}_{-1}[n]$. The reconstruction window is not included, without loss of generality, in the rest of the development. Although, it is worth noting that the appropriate windowing operations must be applied to the original input signal and the synthetic signal being tested in order to obtain a true error measure. The rewritten error term is given by

$$E = \frac{1}{N} \sum_{n=0}^{N-1} (s[n] - g\tilde{s}[n] - \hat{s}_{-1}[n])^2 \quad (5-50)$$

where the N is the length of the analysis frame.

The total error E in equation 5-50 is the term that we want to minimize. The search procedure consists of finding the sub-sampling period P_s that produces a set of

parameters that produces synthetic speech that best matches the input speech in a least square error sense as shown in equation 5-50.

First let us expand equation 5-50 by grouping the input signal and previous frame's synthesized signal together and then computing the square. The total error E computed for each of the candidate sub-sampling periods P_s is given by

$$E^i = \frac{1}{N} \sum_{n=0}^{N-1} (s[n] - \hat{s}_{-1}[n])^2 - 2 \sum_{n=0}^{N-1} g^i \tilde{s}^i[n] (s[n] - \hat{s}_{-1}[n]) + \sum_{n=0}^{N-1} g^{2^i} \tilde{s}^{2^i}[n] \quad (5-51)$$

where the index i selects the i^{th} candidate sub-sampling period P_s^i and its corresponding spectral amplitudes and phases selected by sub-sampling the magnitude spectrum and the phase response with the total error E^i is associated with the current frame's parameters.

As with any minimum error scheme, defining the appropriate match criterion is key to the success of the minimization process. Initially we want to match the original input signal to the corresponding synthetic signal, and in this case we are trying to match the original input time-domain signal. The target, signal to be matched, is defined to be $e^{(0)}[n]$ which is given by

$$e^{(0)}[n] = s[n] - \hat{s}_{-1}[n]. \quad (5-52)$$

The substitution of equation 5-52 into 5-51 leads to the new total error term defined as

$$E^i = \frac{1}{N} \sum_{n=0}^{N-1} e^{(0)2}[n] - \frac{2}{N} \sum_{n=0}^{N-1} g^i \tilde{s}^i[n] e^{(0)}[n] + \frac{1}{N} \sum_{n=0}^{N-1} g^{2^i} \tilde{s}^{2^i}[n]. \quad (5-53)$$

The total error E^i is still dependent on the gain term g^i and the candidate sub-sampling period P_s^i . This is still a complex problem that calls for solving for g^i and P_s^i simultaneously. An alternate approach is to solve for the two parameters sequentially.

The sequential approach is as follows: solve for the optimum g^i using equation 5-53 then solve for P_s^i given the optimum gain g^i .

The gain is found by computing the partial derivative of E^i with respect to the gain g^i and setting equal to zero and solving for the gain g^i . This is given in the following two equations.

$$\frac{\partial E^i}{\partial g^i} = -2 \sum_{n=0}^{N-1} e^{(0)}[n] \tilde{s}^i[n] + 2 \sum_{n=0}^{N-1} g^i \tilde{s}^{2^i}[n] = 0 \quad (5-54)$$

$$g^i = \frac{\sum_{n=0}^{N-1} e^{(0)}[n] \tilde{s}^i[n]}{\sum_{n=0}^{N-1} \tilde{s}^{2^i}[n]} \quad (5-55)$$

Equation 5-55 is the normal form of the cross-correlation. In order to find the optimal minimum MSE sub-sampling period equation 5-53 is set equal to zero as shown by

$$E^i = \sum_{n=0}^{N-1} e^{(0)2}[n] - 2 \sum_{n=0}^{N-1} g^i \tilde{s}^i[n] e^{(0)}[n] + \sum_{n=0}^{N-1} g^{2^i} \tilde{s}^{2^i}[n] = 0. \quad (5-56)$$

The target $e^{(0)}[n]$ is not a function of the index i so equation 5-56 is now rewritten by moving the target energy term to the left side producing

$$\sum_{n=0}^{N-1} e^{(0)2}[n] \geq 2 \sum_{n=0}^{N-1} g^i e^{(0)}[n] \tilde{s}^i[n] - \sum_{n=0}^{N-1} g^{2^i} \tilde{s}^{2^i}[n] \quad (5-57)$$

This inequality is motivated in the following way. The term on the left side of the inequality is the autocorrelation of the target vector. This value represents the best possible match between the target signal $e^{(0)}[n]$ and the synthesized time-domain signal $g^i \tilde{s}^i[n]$. This would suggest that we would want to maximize the term on the right side

of the inequality. This is fine if g^i is a positive value, but the gain g^i is a quantity that is either positive or negative. Since g^i can be negative, there is a possibility that the autocorrelation of the target $e^{(0)}[n]$ is equal to a negative value. This problem is easily solved by the fact that the only way g^i is going to be negative is if the quantity $\sum_{n=0}^{N-1} e^{(0)}[n] \tilde{s}^i[n]$ results in a negative value. If this happens then the term on the right side

of equation 5-57 has a positive result, since g^i and $\sum_{n=0}^{N-1} e^{(0)}[n] \tilde{s}^i[n]$ are both negative. The right side of equation 5-57 is largest when the synthetic speech vector $\tilde{s}^i[n]$ approaches the target signal $e^{(0)}[n]$. This suggests that the optimum minimum MSE is determined by maximizing the quantity $2 \sum_{n=0}^{N-1} g^i e^{(0)}[n] \tilde{s}^i[n] - \sum_{n=0}^{N-1} g^{2i} \tilde{s}^{2i}[n]$ as shown by

$$m_s^i = 2 \sum_{n=0}^{N-1} g^i e^{(0)}[n] \tilde{s}^i[n] - \sum_{n=0}^{N-1} g^{2i} \tilde{s}^{2i}[n]. \quad (5-58)$$

Equation 5-58 is referred to as the *match score* m_s^i for the current set of model parameters and is rewritten in a more compact form by substituting the optimal gain g^i from equation 5-55 into equation 5-58. The result is the following match score

$$m_s^i = \sum_{n=0}^{N-1} g^i e^{(0)}[n] \tilde{s}^i[n] = \frac{\left(\sum_{n=0}^{N-1} e^{(0)}[n] \tilde{s}^i[n] \right)^2}{\sum_{n=0}^{N-1} \tilde{s}^{2i}[n]}. \quad (5-59)$$

This is the squared cross-correlation of the target vector and the synthesized magnitude spectrum normalized by the energy in the synthetic magnitude spectrum corresponding to index i , which directly relates to the optimum sub-sampling period P_s^i .

In summary, a set of candidate sub-sampling periods is selected to represent the current analysis frame. This vector is denoted as P_s^i and typically ranges from 20 samples to 114 samples for speech signals. For each value of P_s^i , a gain g^i and a match score m_s^i are computed as shown in equations 5-55 and 5-59. Since the match score is a maximizing function, the sub-sampling period corresponding to the largest match score is selected to represent the current analysis frame along with the corresponding gain and amplitudes. The following paragraphs discuss the results of the frequency-domain analysis-by-synthesis method derived above.

For ease of development, clarity, and without loss of generality the development of the previous equations are written in terms of vector notation. This is acceptable since this representation is equivalent to dividing the input data into frames. The total error defined in terms of vector notation is

$$\mathbf{E} = \|\mathbf{e}\|^2 \quad (5-60)$$

As stated in the previous section, with any minimum error scheme, defining the appropriate match criterion is key to the success of the minimization process. In this case the match criterion is defined to be the current frame's synthetic speech scaled by a gain \mathbf{g} plus the overlap from the previous frame's synthetic speech. This is in contrast to the frequency-domain approach, which does not consider any past data in the analysis-by-synthesis loop. This match criterion for the time-domain approach is given by

$$\hat{\mathbf{s}} = \mathbf{g}\tilde{\mathbf{s}} + \hat{\mathbf{s}}_{-1}. \quad (5-61)$$

This is the overlap procedure as given in equation 5-20, with the reconstruction window left out for simplicity but without loss of generality. The current frame of

synthetic speech is defined as a gain \mathbf{g} multiplied by the synthetic speech generated using equation 5-10. The addition of the gain term seems appropriate since the Fourier transform is being undersampled. By undersampling, quantization error is introduced into the reconstructed speech; the energy in the reconstructed speech is not equal to the energy in the original input speech as shown in Chapter 4.

The common form for the error vector \mathbf{e} is a perceptually weighted difference between the original input speech vector and the synthetic speech vector defined as

$$\mathbf{e}^i = \mathbf{W}(\mathbf{s} - \hat{\mathbf{s}}^i), \quad (5-62)$$

where \mathbf{W} is a lower triangular matrix that represents the impulse response of the perceptual weighting filter [26]. For the developments in this dissertation \mathbf{W} is set to be the identity matrix \mathbf{I} . The index i determines the synthetic speech vector that corresponds to a given candidate sub-sampling period \mathbf{P}_s and the associated phases and spectral amplitudes.

By substituting equation 5-61 into equation 5-62, a new match criterion is defined as

$$\mathbf{e}^i = \mathbf{W}(\mathbf{s} - \hat{\mathbf{s}}_{-1}) - \mathbf{W}\mathbf{g}^i \tilde{\mathbf{s}}^i \quad (5-63)$$

From this equation we define the target vector as $\mathbf{e}^{(0)} = \mathbf{W}(\mathbf{s} - \hat{\mathbf{s}}_{-1})$. Equation 5-63 is now written compactly in terms of the target vector as

$$\mathbf{e}^i = \mathbf{e}^{(0)} - \mathbf{g}^i \bar{\mathbf{s}}^i \quad (5-64)$$

where $\bar{\mathbf{s}}^i = \mathbf{W}\tilde{\mathbf{s}}^i$.

Substituting equation 5-64 into equation 5-60 leads to the following error metric.

$$\mathbf{E}^i = \|\mathbf{e}^i\|^2 = \mathbf{e}^{(0)\text{T}} \mathbf{e}^{(0)} - 2\mathbf{g}^i \bar{\mathbf{s}}^{i\text{T}} \mathbf{e}^{(0)} + \mathbf{g}^{2i} \bar{\mathbf{s}}^{i\text{T}} \bar{\mathbf{s}}^i \quad (5-65)$$

\mathbf{E} is the total squared error sum corresponding to the candidate sub-sampling period vector \mathbf{P}_s^i and \mathbf{T} is the transpose of the vector. Since \mathbf{E} is a function of both \mathbf{g} and \mathbf{i} then an optimal \mathbf{g} is found for a given index \mathbf{i} . This is accomplished by computing the partial derivative of \mathbf{E} with respect to the gain \mathbf{g} and then setting the derivative equal to zero. This is computationally shown as

$$\frac{\partial \mathbf{E}^i}{\partial \mathbf{g}^i} = -2\bar{\mathbf{s}}^{i\text{T}} \mathbf{e}^{(0)} + 2\mathbf{g}^i \bar{\mathbf{s}}^{i\text{T}} \bar{\mathbf{s}}^i = \mathbf{0}. \quad (5-66)$$

This equation is solved for an optimal gain \mathbf{g} in the minimum mean-squared error sense. The optimal gain \mathbf{g} is found from the normalized cross-correlation between the target vector $\mathbf{e}^{(0)}$ and the synthetic speech vector $\bar{\mathbf{s}}^i$ corresponding to index \mathbf{i} .

$$\mathbf{g}^{(i)} = \frac{\bar{\mathbf{s}}^{i\text{T}} \mathbf{e}^{(0)}}{\bar{\mathbf{s}}^{i\text{T}} \bar{\mathbf{s}}^i} \quad (5-67)$$

To determine the optimal minimum MSE sub-sampling period, equation 5-65 is set equal to zero as shown in the equation below.

$$\mathbf{E}^i = \mathbf{e}^{(0)\text{T}} \mathbf{e}^{(0)} - 2\mathbf{g}^i \bar{\mathbf{s}}^{i\text{T}} \mathbf{e}^{(0)} + \mathbf{g}^{2i} \bar{\mathbf{s}}^{i\text{T}} \bar{\mathbf{s}}^i = \mathbf{0} \quad (5-68)$$

Since the target vector $\mathbf{e}^{(0)}$ is not a function of index \mathbf{i} it is moved to the left side producing

$$\mathbf{e}^{(0)\text{T}} \mathbf{e}^{(0)} \geq 2\mathbf{g}^i \bar{\mathbf{s}}^{i\text{T}} \mathbf{e}^{(0)} - \mathbf{g}^{2i} \bar{\mathbf{s}}^{i\text{T}} \bar{\mathbf{s}}^i. \quad (5-69)$$

The inequality is motivated in the following manner. The term on the left side of the inequality is the autocorrelation (energy) of the target vector $\mathbf{e}^{(0)}$. This value represents the best possible match between the target vector $\mathbf{e}^{(0)}$ and the current frames'

synthetic speech \bar{s}^i . This would suggest that we would want to maximize the term on the right side of the inequality. This approach is fine if g^i is a positive value, but there are no constraints on g^i and it takes on positive and negative values. Since g^i can be negative, there is a possibility that the autocorrelation of $e^{(0)}$ is equal to a negative value. This problem is easily solved by the fact that the only way g^i is negative is if the quantity $\bar{s}^{iT} e^{(0)}$ results in a negative value. If this happens then the term on the right side of equation 5-69 has a positive result, since g^i and $\bar{s}^{iT} e^{(0)}$ are both negative. The right hand side of equation 5-69 is largest when the synthetic speech vector \bar{s}^i approaches the target vector $e^{(0)}$. This suggests that the optimum minimum mean-squared error index i is found by maximizing the quantity $2g^i \bar{s}^{iT} e^{(0)} - g^{2i} \bar{s}^{iT} \bar{s}^i$, as shown in equation 5-70.

$$m_s^i = 2g^i \bar{s}^{iT} e^{(0)} - g^{2i} \bar{s}^{iT} \bar{s}^i \quad (5-70)$$

Equation 5-70 is referred to as the match score for the current set of model parameters and is rewritten by substituting the optimal gain g from equation 5-67 into equation 5-70. The result is the following match score

$$m_s^i = g^i \bar{s}^{iT} e^{(0)} = \frac{(\bar{s}^{iT} e^{(0)})^2}{\bar{s}^{iT} \bar{s}^i}. \quad (5-71)$$

This is the squared cross-correlation of the target vector and the synthesized speech vector normalized by the energy in the synthetic speech vector corresponding to index i .

In summary, a set of sub-sampling periods is selected as candidates for representing the current analysis frame. This vector is denoted as P_s^i and typically ranges from 20 to 114 for speech signals. For each value of P_s^i , a gain g^i and a match score m_s^i

are computed as shown in equations 5-67 and 5-70. Since, the match score is a maximizing function, the sampling period corresponding to the largest match score is selected to represent the current analysis frame along with the corresponding gain, amplitudes, and phases.

5.5.2 Simulation

5.4.2.0 Introduction

This section describes the simulation of the time-domain analysis-by-synthesis method derived in this section. This simulation, just like the frequency-domain approach, is also written in the 'C' programming language on a Sun Sparc Workstation Ultra 170. The idea again is to prove the concept of the time-domain analysis-by-synthesis method so complexity is considered to be of secondary importance. In contrast to the low bit rate nature of the frequency-domain analysis-by-synthesis method, the time-domain analysis-by-synthesis method is naturally targeted at high bits rates (approximately 13,000 bps and up).

The input signals used in this simulation, as in the frequency-domain method, are quantized using 16 bits and a sampling frequency of 8,000 samples per second. The input signal is windowed using a 240 (30ms) point square-root of Hamming window same as in the frequency-domain simulation. This window is used to compute the magnitude spectrum and phase response for each frame. The reconstruction window is a 240 point triangle window, other windows such as Hamming, hanning, and rectangle are possible alternate windows. The analysis window is updated by shifting in 7.5ms (60 samples)

intervals. The center of the analysis window is the time reference, which results in an overlap of the 90 samples in the past and 90 samples in the future.

The magnitude spectrum of the input signal is computed using the DFT. The length is chosen to provide the appropriate resolution for the selection of the spectral amplitudes and their corresponding phases. As in the frequency-domain simulation, the DFT length chosen is $M = 16,384$. The resulting magnitude spectrum and phase response is then sub-sampled producing a set of spectral amplitudes and corresponding phases for a given candidate sub-sampling period. The phase information is found using the sub-sampling formulas of section 5.2.

The sub-sampling period P_s is selected to fall in the range of 20 samples to 114 samples. In contrast to the frequency-domain approach, a single stage process is used to determine the sub-sampling period for the current frame. The sub-sampling range is linearly quantized over the sub-sampling range using 8 bits, which results in 256 possible candidate sub-sampling periods to test. This approach is chosen over the two stage frequency-domain approach because the time-domain method is lower complexity so is afforded the luxury of a more exhaustive search.

In contrast to the frequency-domain analysis-by-synthesis method, no voicing decisions are necessary. This results from the fact that the phase contains the voicing information. This has the advantage of being more robust and less susceptible to errors especially in the transition regions. The disadvantage is that the bit rate increases to accommodate the extra information.

For each of the candidate sub-sampling periods a synthetic signal is generated in the time-domain using the sinusoidal model with OLA. These synthetic signals are

compared in a mean-squared error sense in the time-domain to the original time-domain input signal. An optimum gain and a match score are found for each of the corresponding synthetic signals where the parameter set producing the highest match score is chosen to represent the current frame.

The following paragraphs discuss in more detail the signals used in testing the concept of time-domain analysis-by-synthesis, the sub-sampling process, the analysis-by-synthesis loop, the match scores, and the resulting sub-sampling contour.

5.5.2.1 Test Signals

The same three signals used in testing the frequency-domain analysis-by-synthesis method are used to the time-domain analysis-by-synthesis method. The three signals are an all voiced signal, an all unvoiced signal, and a real speech signal containing the word “Figure”. These signals are presented in Figures 5-5, 5-6, and 5-7.

5.5.2.2 Sub-Sampling Process

The sub-sampling process is the same as the sub-sampling process described in section 5.4.2.2. The main difference between the frequency-domain method and the time-domain method is the addition of sub-sampling the phase response for the time-domain method. The phase response is sub-sampled at exactly the same points as the magnitude spectrum. These points are determined using the methods presented in the analysis section of this chapter.

5.5.2.3 Analysis-By-Synthesis Loop

This section describes in more detail the operation of the time-domain analysis-by-synthesis loop. A comparison of the appropriately windowed original input time-

domain signal with the appropriately windowed synthetic signal is provided for a particular frame for all three tests signal and multiple sub-sampling periods.

The first test signal is the all voiced constant tone. Figure 5-45 displays the windowed input signal, the solid line, with the windowed synthetic signal, the dashed line. The sub-sampling period for this case is $P_s = 20$. As in the case of the frequency-domain method the low sub-sampling rate does not produce a good match to the input time-domain signal.

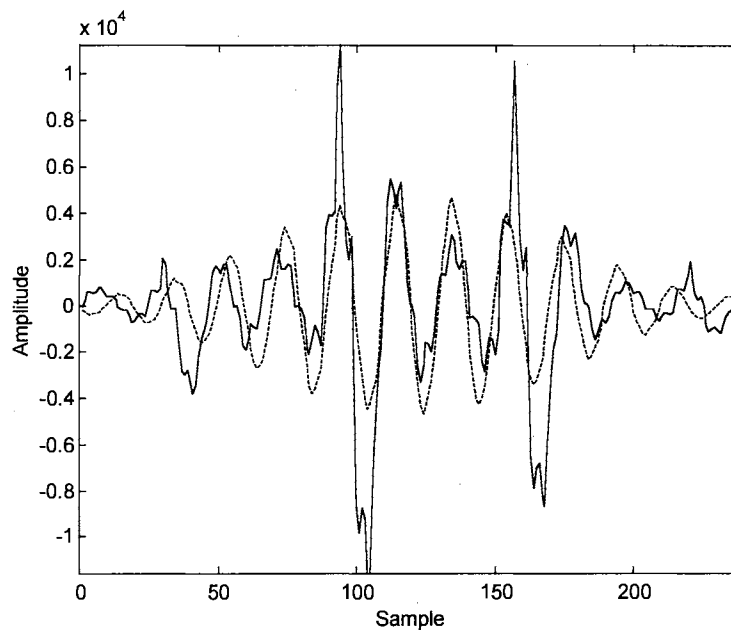


Figure 5-45. Original Input Signal and All Voiced Synthetic Signal for $P_s = 20$

Figure 5-46 shows the windowed input signal with the windowed synthetic signal at a sub-sampling period of $P_s = 114$. For this case, the synthetic signal does match a couple of the peaks in the time-domain signal but mostly varies away from the input signal. Note, based on this observation the best match is not obtained by artificially sub-sampling at a high rate.

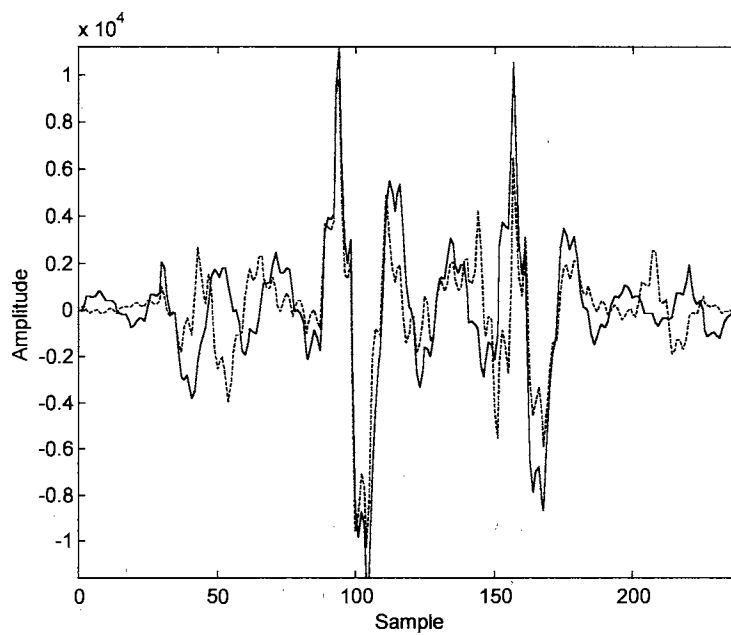


Figure 5-46. Original Input Signal and All Voiced Synthetic Signal for $P_s = 114$

Figures 5-47, 5-48, and 5-49 show how the synthetic signals converge to the input signal as the sub-sampling period producing the highest match score is approached. In all three cases the high amplitude regions are matched quite well. The low amplitude region is where most of the error between the signals is associated. In Figures 5-47 and 5-49 the synthetic signal does not match as closely as the synthetic signal in Figure 5-48 in the range of the 200th sample and up.

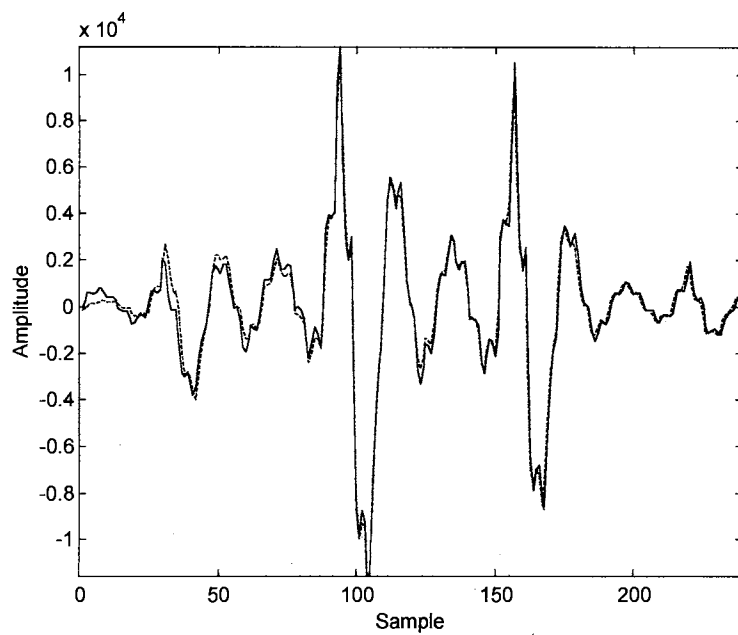


Figure 5-47. Original Input Signal and All Voiced Synthetic Signal for $P_s = 62.76$

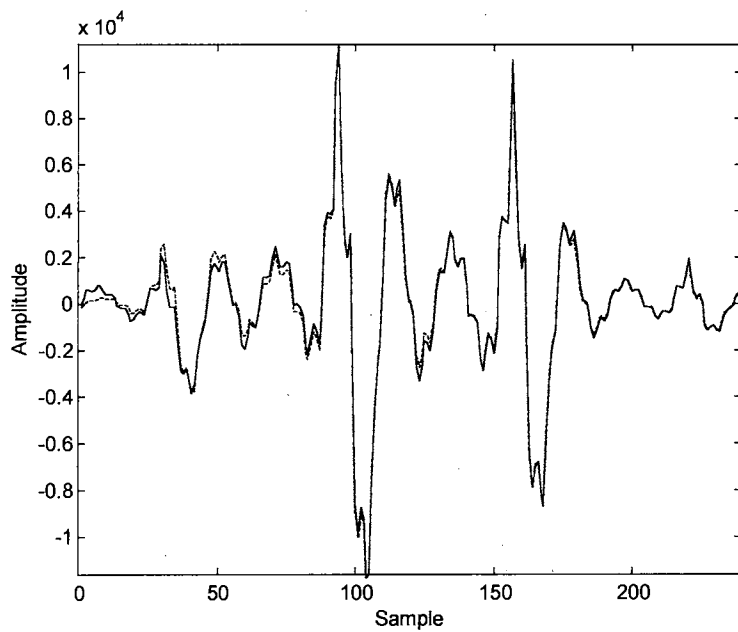


Figure 5-48. Original Input Signal and All Voiced Synthetic Signal for $P_s = 63.13$

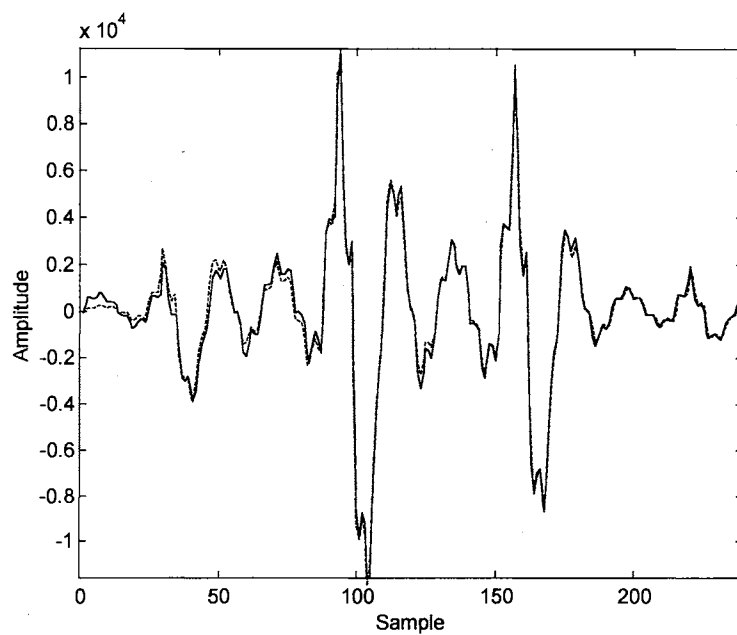


Figure 5-49. Original Input Signal and All Voiced Synthetic Signal for $P_s = 63.50$

The next three figures show the relationship between the original all unvoiced time-domain signal and the all unvoiced synthetic signal for the sub-sampling periods of $P_s = 20$, $P_s = 114$, and $P_s = 74.92$. The windowed synthetic signals are presented in Figures 5-50, 5-51, and 5-52, respectively. Once again the lowest sub-sampling period, shown in Figure 5-50, does not provide a good match to the original windowed input signal. The synthetic signal of Figure 5-51 appears to be varying in a pattern that is similar to the original time-domain input signal. The third synthetic signal, shown in Figure 5-52, also appears to be varying in a pattern similar to the original time-domain input signal. This supports the fact that it takes a large number of sinusoids to generate synthetic unvoiced signals. This is the same observation made in the frequency-domain analysis-by-synthesis method.

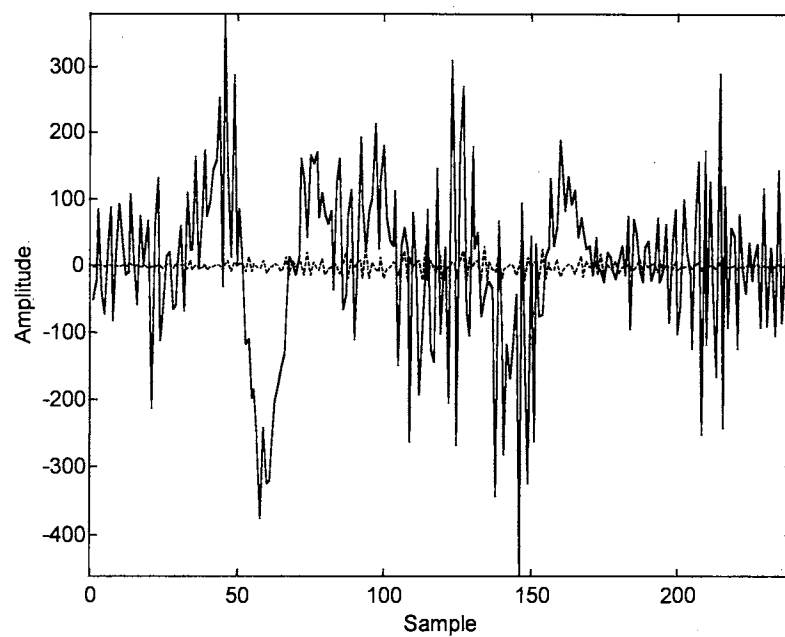


Figure 5-50. Original Input Signal and All Unvoiced Synthetic Signal for $P_s = 20$

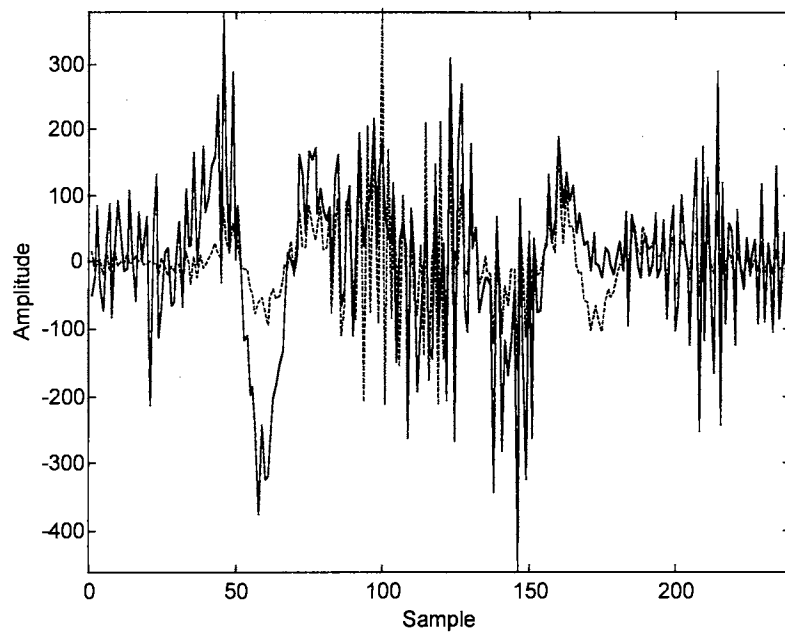


Figure 5-51. Original Input Signal and All Unvoiced Synthetic Signal for $P_s = 114$

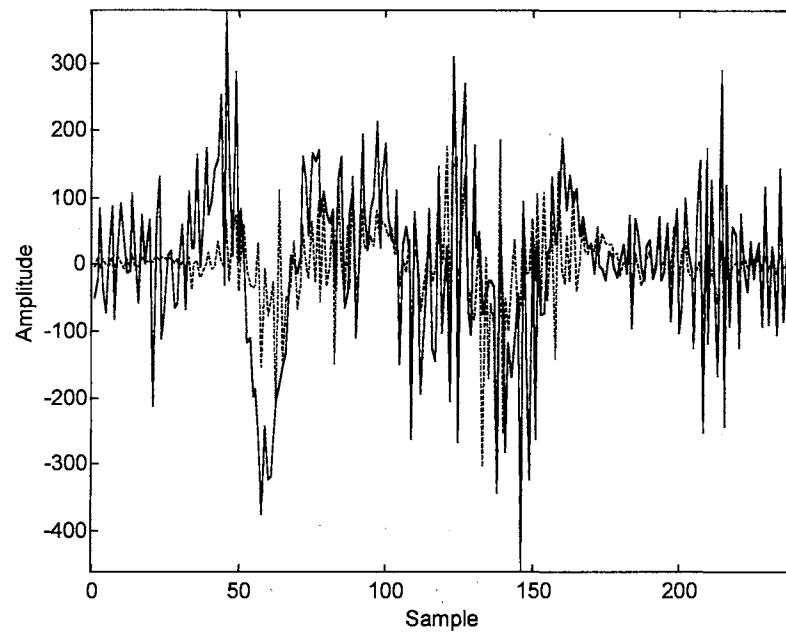


Figure 5-52. Original Input Signal and All Unvoiced Synthetic Signal for $P_s = 74.92$

The following set of figures is an analysis of the response of the time-domain analysis-by-synthesis to a frame of a real speech signal. Figure 5-53 is the synthetic signal for the sub-sampling period $P_s = 20$. The low number of sample points does not allow this sub-sampling period to model accurately the high energy regions of the input time-domain signal. In Figure 5-54, the synthetic signal corresponding to the sub-sampling period of $P_s = 114$ is provided. The large number of sample points allows this synthetic signal to model the input time-domain signal with more accuracy than the sub-sampling period of $P_s = 20$, at least in the high energy regions. The low energy regions are still not modeled as accurately as they should be.

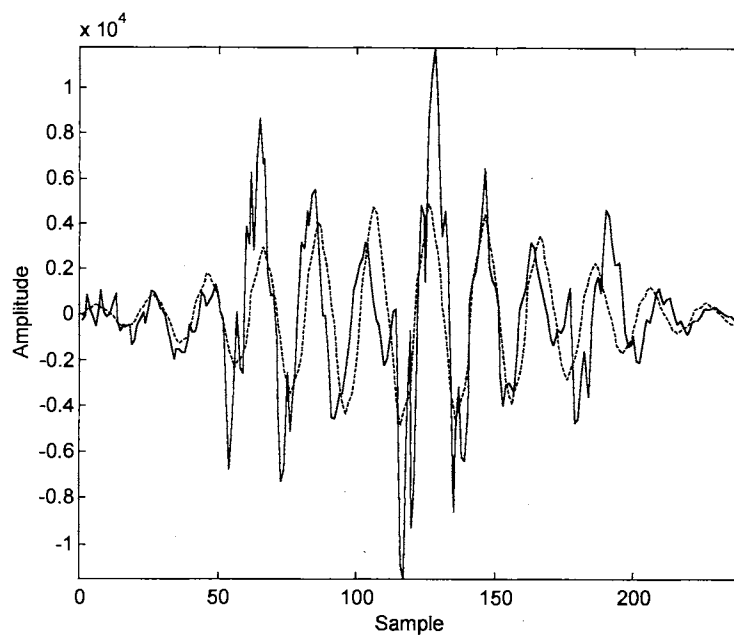


Figure 5-53. Original Input Signal and “Figure” Synthetic Signal for $P_s = 20$

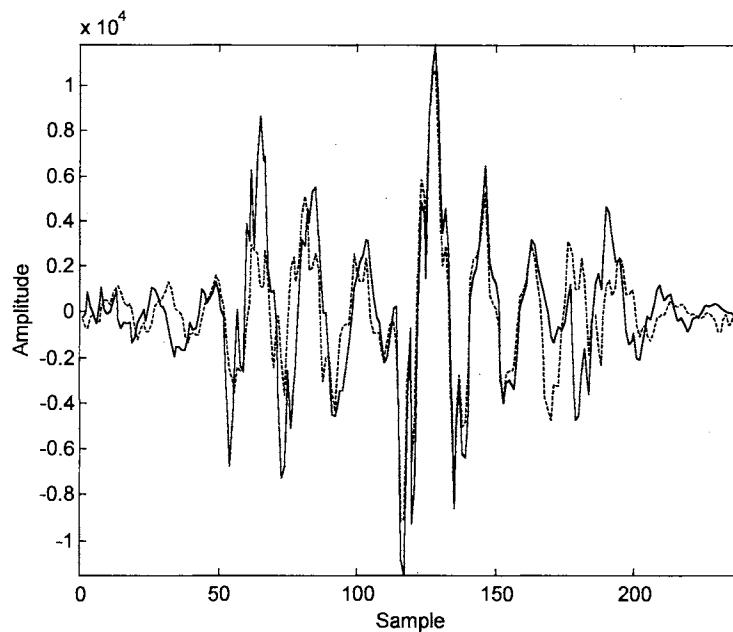


Figure 5-54. Original Input Signal and “Figure” Synthetic Signal for $P_s = 114$

In the frequency-domain method, the selection of the best synthetic magnitude spectrum through observation is not an easy task. This observation is also true in the time-domain. As noted in the frequency-domain section this particular frame of speech

contains both voiced and unvoiced excitation. The three synthetic signals corresponding to the sub-sampling periods $P_s = 62.4$, $P_s = 62.76$, and $P_s = 63.12$ are presented in Figures 5-55, 5-56, and 5-57. In all three cases, the synthetic signal seems to closely match the original time-domain input signal. The high energy region is modeled well and the error between the original signal and the synthetic signal stems from the low energy regions.

The next section discusses the match scores that are associated with the test signals used for analysis in this section.

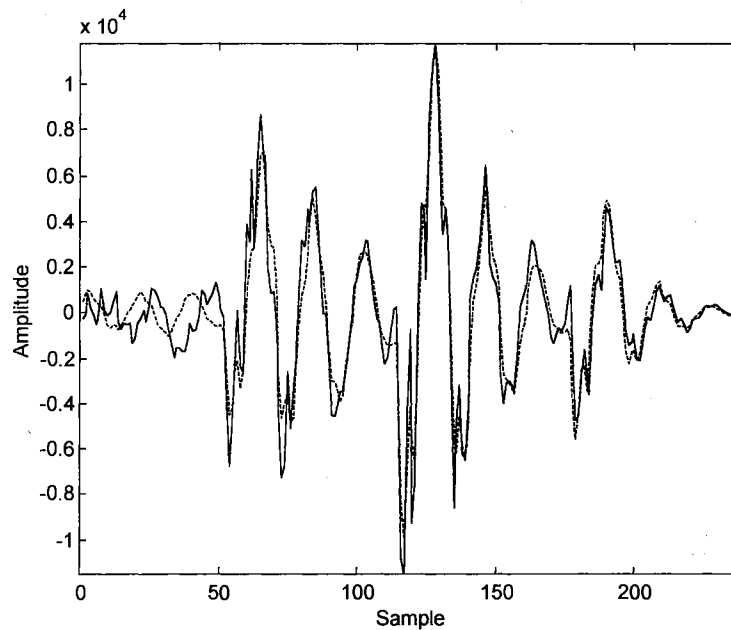


Figure 5-55. Original Input Signal and “Figure” Synthetic Signal for $P_s = 62.40$

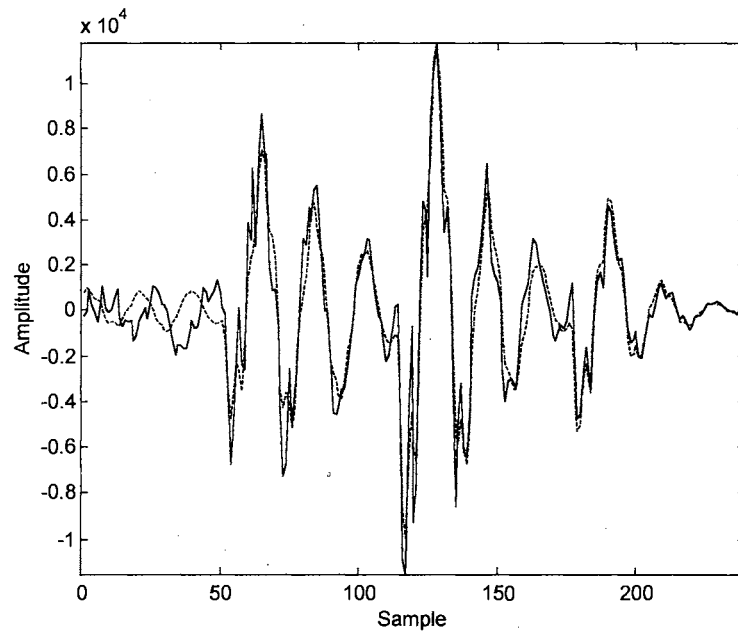


Figure 5-56. Original Input Signal and “Figure” Synthetic Signal for $P_s = 62.76$

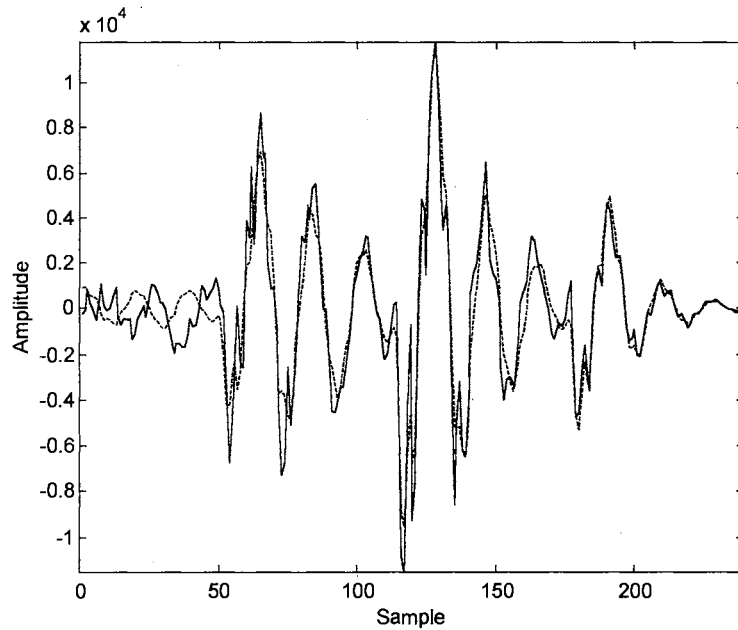


Figure 5-57. Original Input Signal and “Figure” Synthetic Signal for $P_s = 63.12$

5.5.2.4 Match Scores

This section discusses the match scores for the frame of speech that is being analyzed. The match scores represent how well the windowed synthetic time-domain

signal, for each candidate sub-sampling period given an optimum gain, corresponds to the original windowed input time-domain signal.

The process of searching for the optimum sub-sampling period is found by a more exhaustive search than the frequency-domain analysis-by-synthesis method presented in a previous section. The sub-sampling range is quantized to 256 levels and each one of the candidates is searched. The candidate sub-sampling period producing the highest match score is chosen to represent the current frame along with its corresponding spectral amplitude and phases.

The match scores for the synthetic all voiced time-domain signal are shown in Figure 5-58. Again, it is worth noting that a number of methods exist for finding the minimum solution to the mean-squared problem developed in section 5.5, but for the methods to work properly only one minimum should exist. Figures 5-58, 5-59, and 5-60 show that for speech or speech like signals several local minimum and maximum exist. So a more exhaustive type of search is necessary to find the optimal solutions to the mean-squared error problem.

The all voiced signal has a fundamental frequency equal to approximately 63.4 samples. The sub-sampling period producing the highest match score, given the OLA sinusoidal model of reconstruction, is 63.12 samples. Another method of reconstruction in the analysis-by-synthesis loop may produce a slightly different result. Once again the effect of pitch doubling or halving does not appear to be a problem although the sub-sampling period corresponding to a pitch doubling does produce a maximum.

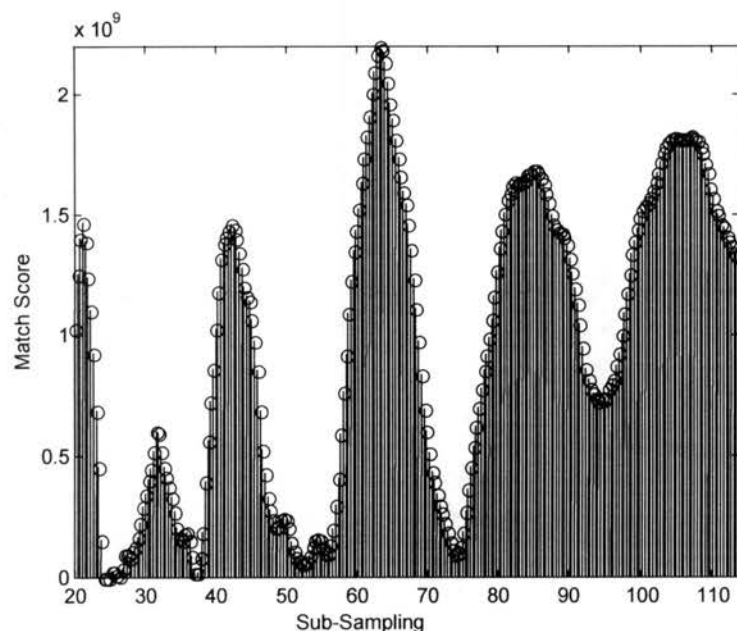


Figure 5-58. Match Scores for Sub-Sampling Periods of the All Voiced Signal

The match scores corresponding to the windowed synthetic all unvoiced signal are presented in Figure 5-59. As stated in previous sections, the synthetic all voiced signal produces a closer match to the original time-domain signal as the sub-sampling period is increased. This fact is not quite as evident as in the frequency-domain method. The match scores do increase as the sub-sampling is increased but the use of the phase information shows a sub-sampling period that best fits the original time-domain signal using the sinusoidal model of reconstruction.

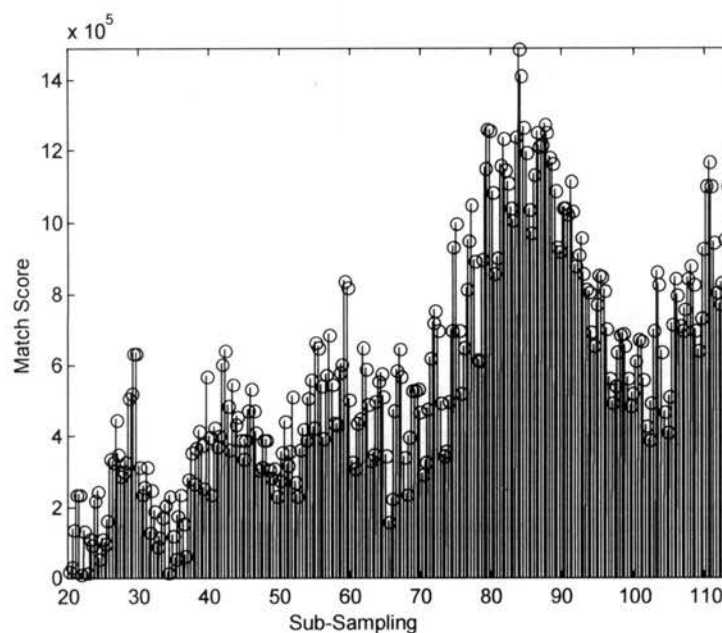


Figure 5-59. Match Scores for Sub-Sampling Periods of the All Unvoiced Signal

As in the case of the frequency-domain approach, the match score contours produced in the first two signals result in the appropriate selection of the sub-sampling period and the corresponding spectral amplitudes and phases for the nearly ideal conditions. The final and most important test is to determine the match scores in response to a frame of real speech, the word “Figure”. Similar to the all voiced signal, there is more than one possible maximum, but the highest match score corresponds to the sub-sampling period of 62.76. This is a slightly different result than the frequency-domain analysis-by-synthesis approach, which produced a sub-sampling period of 114. The next section discusses the resulting sub-sampling period contour produced after analyzing each of the test signals entirely.

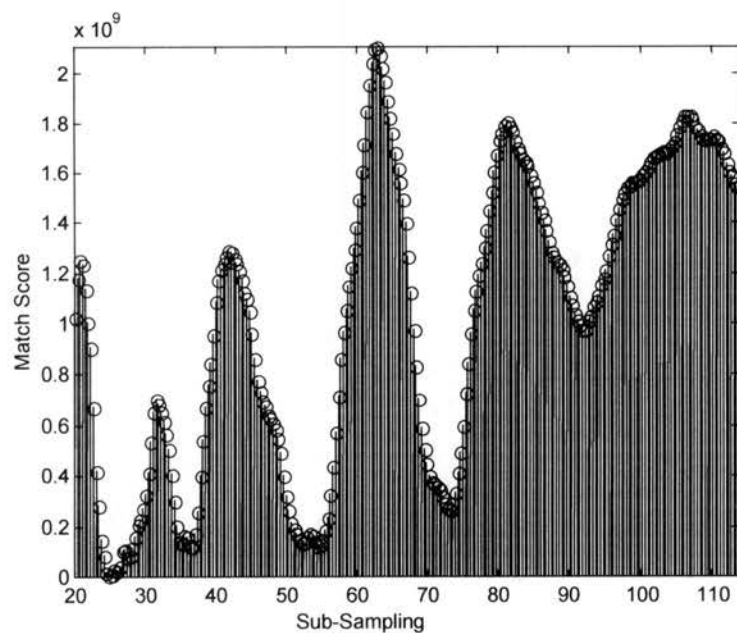


Figure 5-60. Match Scores for Sub-Sampling Periods of the Word “Figure”

5.5.2.5 Sub-Sampling Period Contour

This section discusses the results of the sub-sampling period contour obtained by analyzing the three test signals, all voiced, all unvoiced, and the word “Figure”. As noted previously, the all voiced signal is a constant tone so the sub-sampling period contour is expected to also be constant. Figure 5-61 shows the sub-sampling period contour produced using the time-domain analysis-by-synthesis method to estimate the sub-sampling period. The contour as expected is constant except in the transition regions at the beginning and the end of the signal.

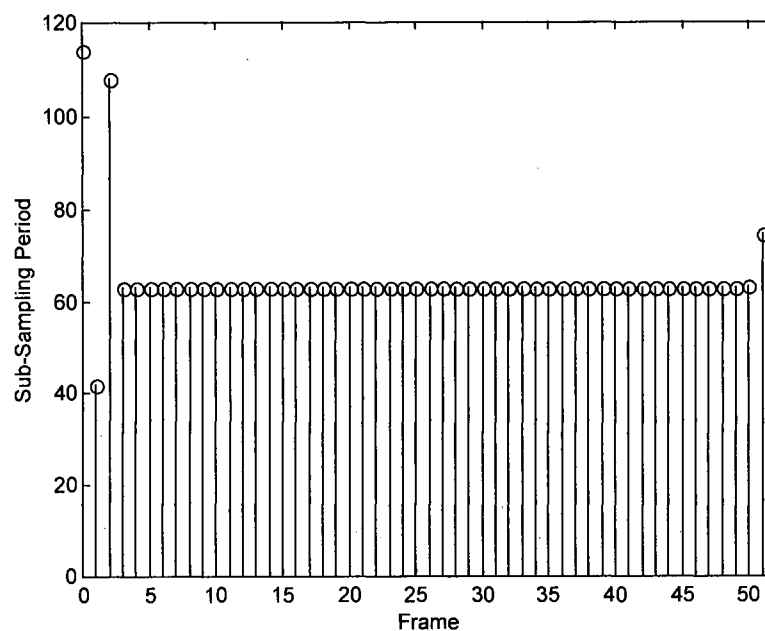


Figure 5-61. All Voiced Sub-Sampling Period Contour

The next test signal is the all unvoiced signal. In contrast to the all voiced signal it is expected that the contour will not be constant but will be biased towards the high sub-sampling periods. The sub-sampling period contour determined using the time-domain analysis-by-synthesis method for the all unvoiced signal is shown in Figure 5-62. In Figure 5-62 the sub-sampling period contour does vary from frame-to-frame and in general is biased to the higher sub-sampling periods.

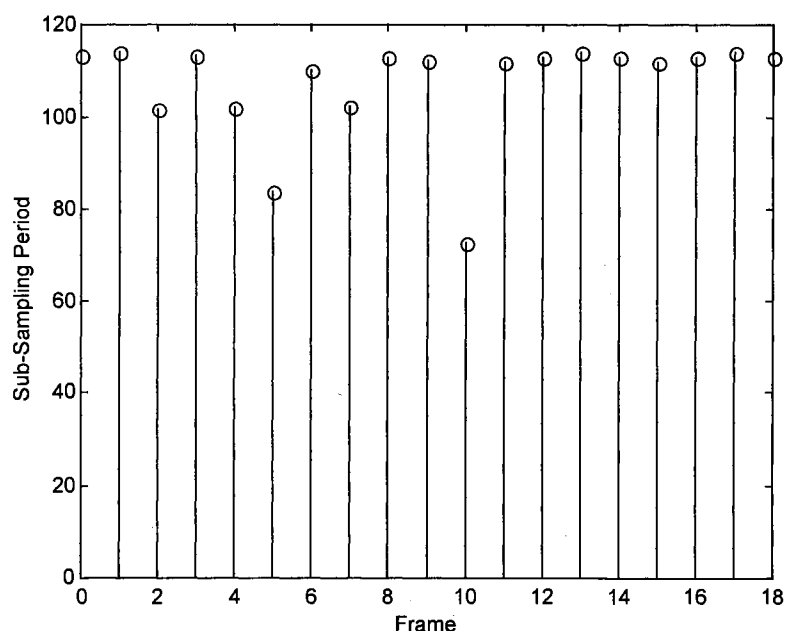


Figure 5-62. All Unvoiced Sub-Sampling Period Contour

The third test signal and the real key to the success of the time-domain analysis-by-synthesis method, as in the frequency-domain method, is a real speech signal. The real speech signal has time-varying properties that were not present in either the all voiced or all unvoiced signal. The sub-sampling period contour is expected to vary slightly from frame-to-frame during the voiced and unvoiced regions but should be biased towards the higher sub-sampling periods in the unvoiced regions. Figure 5-63 presents the sub-sampling period contour for the real speech signal produced using the time-domain analysis-by-synthesis method to estimate the sub-sampling period for each frame. The contour as expected is not constant at any time but varies from frame-to-frame. In the voiced regions the sub-sampling period contour is smoothly varying and in the unvoiced regions the sub-sampling period contour is biased towards the higher sub-sampling periods. The next section describes the synthetic signals produced from the parameter estimates obtained using the time-domain analysis-by-synthesis method.

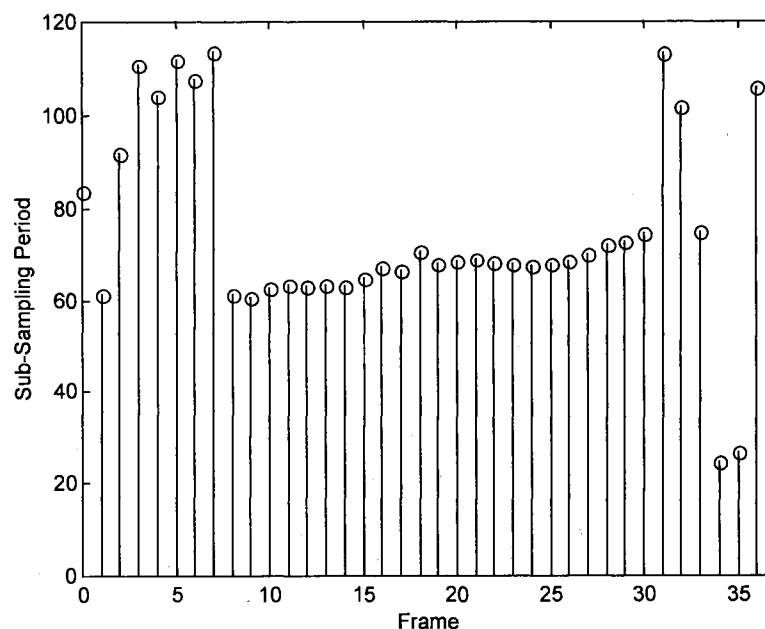


Figure 5-63. “Figure” Sub-Sampling Period Contour

5.5.2.6 Synthesized Test Signals

This section looks at the resulting synthetic signals produced from the parameter estimates obtained from the time-domain analysis-by-synthesis approach and using the sinusoidal model for reconstructing the synthetic signal.

The three test signals, all voiced, all unvoiced, and the word “Figure” are shown in Figures 5-64, 5-65, and 5-66, respectively. The synthetic all voiced signal is almost an exact replica of the original time-domain signal, other than being delayed. The main difference is in the amplitude, which is a result of the quantization of the spectral amplitudes and phases. The onset, while not exact, is much sharper than the frequency-domain method. In contrast to the frequency-domain method the time-domain analysis-by-synthesis method produces a synthetic signal that is in phase with the original.

The synthetic all unvoiced signal is a good approximation to the original time-domain all unvoiced signal, other than being slightly delayed. Because of the inclusion of

the phase information the synthetic unvoiced signal is a more realistic match than the frequency-domain method.

The reconstructed signal for the word “Figure” does resemble the original time-domain signal, besides being delayed. The synthetic version has slightly smoother amplitude variations but the onset is modeled much closer than with the frequency-domain method. The result is that the time-domain analysis-by-synthesis method produces synthetic signals that appear to match the original time-domain signals with minimum error.

In the end the true test is the listening test. The signals synthesized using the time-domain analysis-by-synthesis produced reconstructed signals that in most cases are indistinguishable from the original.

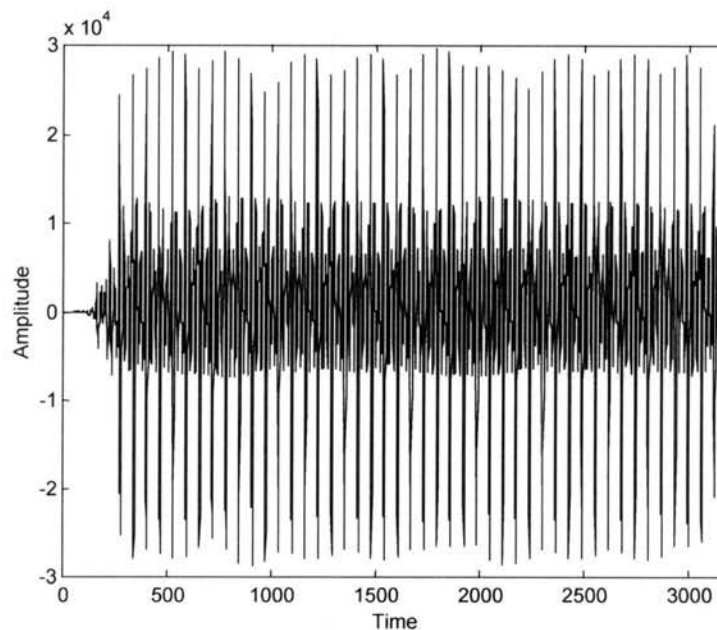


Figure 5-64. Synthetic All Voiced Signal

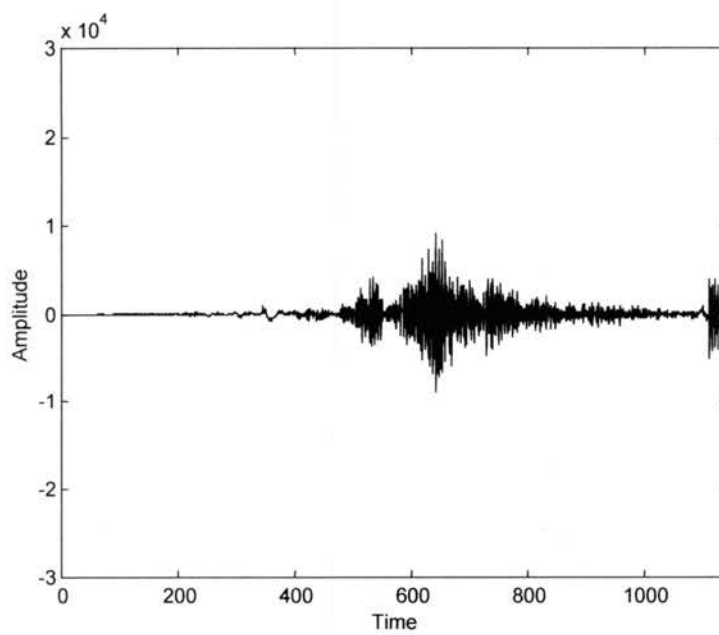


Figure 5-65. Synthetic All Unvoiced Signal

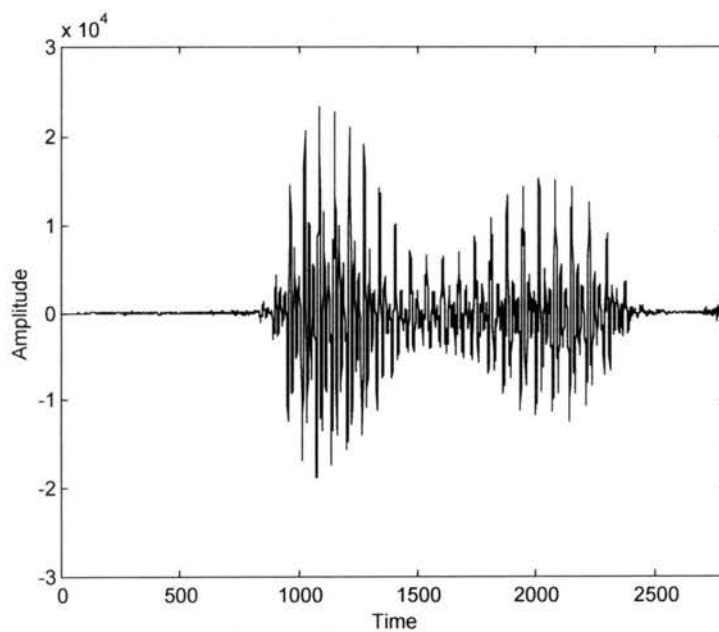


Figure 5-66. Synthetic Word "Figure"

5.5.2.7 Conclusion

In section 5.5 a time-domain analysis-by-synthesis method of selecting the appropriate parameters for the sinusoidal model using OLA is developed. A complete theoretical development is presented along with the simulation results.

The simulation is tested on three different test signals, an all voiced, an all unvoiced, and a speech signal (the word “Figure”). The all voiced signal is used to test the response of the time-domain method using a constant tone. The all unvoiced signal is used to test the response to the time-domain method using a noise signal. Both of these signals represent the ideal conditions for pure voiced and pure unvoiced speech. The word “Figure” is added to test the response of the time-domain method of a more realistic signal.

The time-domain analysis-by-synthesis method is shown to respond as expected given all three of the test signals. While the complexity of the analysis-by-synthesis loop of the time-domain method is much less than the frequency-domain method, a trade is made for shorter correlation and a more exhaustive search. The frequency-domain method required correlation on the length of 8192 but the time-domain method only needs correlation on lengths the size of the frame, 240 points in this simulation.

The frequency-domain sub-sampling search range consisted of 95 integer searches and at most 21 fractional searches, which totals 116 searches. The time-domain method quantizes the sub-sampling range using 8 bits so the total number of searches is 256, more than 2 times that of the frequency-domain method.

Ignoring complexity, the time-domain analysis-by-synthesis has a major advantage over the frequency-domain method because the time-domain method does not

require any voicing decisions. The voicing decisions as noted earlier may handicap the frequency-domain method. The drawback to no voicing decisions is the higher bit rate required to code the phase information.

The next chapter discusses a mid/low bit rate vocoder that combines the best of the frequency-domain and time-domain methods. This vocoder is an analysis-by-synthesis vocoder using the sinusoidal model with OLA for reconstruction.

6 SINUSOIDAL MODEL ANALYSIS-BY- SYNTHESIS VOCODER

6.0 Introduction

This chapter describes the implementation of an analysis-by-synthesis sinusoidal vocoder based on a combination of the frequency-domain and time-domain analysis-by-synthesis methods developed in Chapter 5. This new vocoder, referred to as SMABS, is targeted for 8,000 bps (1 bit per sample for speech sampled at 8,000 samples per second).

In Chapter 5 of this dissertation two novel methods for determining the model parameters for the sinusoidal model of reconstruction using analysis-by-synthesis are presented. One method is developed in the frequency-domain using a no phase assumption and the other method is developed in the time-domain based on the assumption that phase information is available.

The SMABS vocoder uses the time-domain analysis-by-synthesis method for estimating the model parameters. The time-domain analysis-by-synthesis method is chosen over the frequency-domain analysis-by-synthesis because of the computational complexity of the frequency-domain approach. This topic is discussed further in the future research section of Chapter 7.

The disadvantage of the time-domain approach is the high bit rate. For this reason, the time-domain approach is used to perform the parameter estimation but voicing

decisions are transmitted instead of the phase information. This results in a considerable reduction in the bit rate. Since the phase information is lost in transmission it is regenerated in the synthesizer using the techniques discussed in the frequency-domain approach. As a side result, a new gain term is needed because of the loss of phase information.

The model parameters are coded into an 8,000 bps bit stream. These parameters are decoded and synthetic speech is reconstructed using a sinusoidal model with OLA. The phase information, as stated, is generated in the synthesizer using a linear phase model along with a system phase component that is determined from the spectral envelope.

The following sections describe the procedures used to estimate, quantize, and code the relevant parameters into an 8,000 bps bit stream, decode the coded parameters from the bit stream, and reconstruct high quality speech from the estimated parameters. It is assumed that the reader is familiar with short-time analysis, so the details of the implementation are not presented.

6.1 Analyzer

6.1.0 Introduction

Speech analysis is performed sequentially every 30ms on overlapping analysis frames, producing a frame rate of approximately 33 analysis frames/second. Each analysis frame is then split into four 7.5ms subframes. An alternating superframe/subframe analysis strategy is applied so as to reduce the total number of parameters being produced

each second, thus reducing the required bit rate. Each superframe consists of a full update of all coder parameters, while each subframe represents only a partial update. A full analysis and update occurs twice for each analysis frame. The partial updates also occur twice for each analysis frame. This framing strategy is found to be sufficient for good temporal resolution.

Analysis consists of prefiltering, parameter estimation, quantization, and coding. Parameter decoding and frame-by-frame reconstruction of the coded speech form the synthesis stage. The relevant parameters, which are used to represent the input speech waveform, are sub-sampling period, vocal tract spectrum, voicing decisions, and a gain. A block diagram of the analyzer is shown in Figure 6-1.

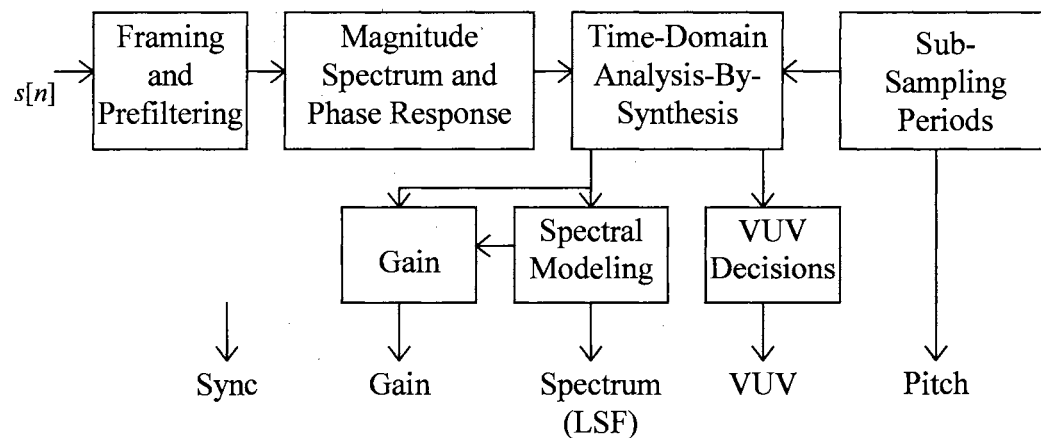


Figure 6-1. Block Diagram of SMABS Analyzer

The SMABS analyzer estimates the following parameters: sub-sampling period (pitch), voicing, spectrum, and gain. These parameters are quantized and coded for either transmission or storage. The input speech is framed, filtered and windowed into multiple data paths. The pitch estimate is determined using time-domain analysis-by-synthesis and the voicing decisions are computed as a result of the best pitch estimate determined by the

analysis-by-synthesis. The spectral amplitudes are modeled using linear prediction and a gain is found by equalizing the energy in the original spectrum and the synthetic spectrum.

6.1.1 Pre-Filtering and Windowing

The input speech $s[n]$ is filtered with a high pass filter with a cutoff frequency of approximately 70 Hz. This filter is used mainly for removing the low frequency components that may inhibit the parameter estimation. For example, the pitch is only estimated over the range 70 Hz to 400 Hz so frequencies below 70 Hz are not needed for analysis.

The high pass filter is a 5th order elliptic filter with 0.25 dB of ripple in the passband and more than 20 dB of attenuation at 60 Hz. The frequency response of this high pass filter is pictured in Figure 3-3 and the filter transfer function, with quantized coefficients, is provided in equations 3-1 and 3-2.

The high pass filtered signal, $s_{HPF}[n]$, is computed using

$$s_{HPF}[n] = \sum_{r=0}^{N-1} s[r] h_{HPF}[n-r] \quad (6-1)$$

where N is the length of the data segment, $s[n]$ is the input speech signal and $h_{HPF}[n]$ is the impulse response of the transfer function $H_{HPF}(z)$ given in equation 3-1.

After the input speech is filtered, it is windowed using a rectangular window, square-root of Hamming window, Hamming window, and a triangular window. The windowing operations are given by

$$s_V[n] = s_{HPF}[n] w_R[n] \quad (6-2)$$

$$s_M[n] = s_{HPF}[n]w_{SQRTH}[n] \quad (6-3)$$

$$s_s[n] = s_{HPF}[n]w_H[n] \quad (6-4)$$

$$s_T[n] = s_{HPF}[n]w_T[n] \quad (6-5)$$

where $s_v[n]$ represents the data used to aid in determining voicing decisions, $s_M[n]$ represents the data used to compute the magnitude spectrum for sub-sampling, $s_s[n]$ represents the data used for computing the linear prediction coefficients, and $s_T[n]$ represents the data used to determine the target data for the analysis-by-synthesis loop.

These windows are defined by

$$w_R[n] = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & \text{otw} \end{cases} \quad (6-6)$$

$$w_{SQRTH}[n] = \begin{cases} \left(0.54 - 0.46 \cos\left[\frac{2\pi n}{N-1}\right]\right)^{\frac{1}{2}} & 0 \leq n \leq N-1 \\ 0 & \text{otw} \end{cases} \quad (6-7)$$

$$w_H[n] = \begin{cases} 0.54 - 0.46 \cos\left[\frac{2\pi n}{N-1}\right] & 0 \leq n \leq N-1 \\ 0 & \text{otw} \end{cases} \quad (6-8)$$

$$w_T[n] = \begin{cases} \frac{2n}{N} & 0 \leq n \leq \frac{N}{2} \\ 2 - \frac{2n}{N} & \frac{N}{2} < n < N \\ 0 & \text{otw} \end{cases} \quad (6-9)$$

where N is defined by the length of the analysis window and the triangular window is scaled appropriately so that when fully overlapped sums to 1. For an analysis window of 240 points with a 60 sample update (an overlap of four) the scaling factor is one-half.

6.1.2 Pitch Estimate

The pitch is estimated using the time-domain analysis-by-synthesis technique developed in Chapter 5. First the magnitude spectrum and phase response are computed using equations 5-1 through 5-4 with a DFT length of 16,384. The magnitude spectrum and phase response are sub-sampled to produce an estimate for the spectral amplitudes and corresponding phases. A set of spectral amplitudes and phases is found for each candidate sub-sampling period. The candidate sub-sampling period ranges from 20 samples to 114 samples. The sub-sampling range is constrained to only 256 values using a linear spacing, which allows this parameter to be coded using 8 bits.

Each set of spectral amplitudes and phases is applied to the sinusoidal model of reconstruction to obtain an estimate for the current frame of speech. The synthetic signal is then compared to the original in a mean-squared error sense. The sub-sampling period producing the best set of spectral amplitudes and phases is determined by equations 5-55 and 5-58.

6.1.3 Voicing

The voiced and unvoiced decisions are the heart of any MBE based analysis model. It is assumed that the speech spectrum is composed of both voiced and unvoiced bands, thus multiple voicing decisions are made in each frame. This is equivalent to considering the excitation to contain both periodic and aperiodic components simultaneously in the same frame. The MBE approach has been shown to produce higher quality synthetic speech and is more robust than the single voicing decision approach. For this reason the multiple voicing decision approach is used in this vocoder, since the phase

information is not being transmitted. This vocoder uses a non-linear band structure, containing 4 bands, as defined in Chapter 3 and reproduced here for clarity

$$B_L = \begin{cases} 5 & 11 & 15 & 16-25 & L \geq 42 \\ 5 & 9 & 13 & 5-14 & 32 \leq L < 42 \\ 5 & 7 & 7 & 6-12 & 25 \leq L < 32 \\ 3 & 5 & 5 & 2-11 & 15 \leq L < 25 \\ 3 & 3 & 3 & 0-6 & L < 15 \end{cases} \quad (6-10)$$

where L represents the number of harmonics corresponding to the current frame.

The estimation of the voicing decisions turned out to be more difficult than expected. The frequency-domain method described in the EMBE vocoder was tested first. The problem with the frequency-domain method of estimating the voicing is that the sub-sampling period determined using the time-domain analysis-by-synthesis method is not compatible thus resulting in very poor voicing decisions. For this reason an alternate voicing scheme is needed. The method developed for the SMABS vocoder is described in the following paragraphs.

One possible solution for estimating the voicing decisions is to use the signal-to-noise (SNR) ratio. This seems reasonable since during voiced periods the match between the original time-domain signal and the synthetic time-domain is high indicating a small mean-squared error. During periods of unvoiced speech the SNR is low indicating that the mean-squared error is high even with the optimum sub-sampling period. In the transition regions the SNR falls somewhere in between the voiced and unvoiced values.

Based on experience and observation the voicing decisions need to vary in a smooth fashion from frame-to-frame. For this reason a predetermined set of voicing combinations is selected. The possible combinations of voicing decisions are given by

$$V_{B_L} = \begin{cases} [1 & 1 & 1 & 1] & SNR > 35 \\ [1 & 1 & 1 & 0] & 28 < SNR < 35 \\ [1 & 1 & 0 & 0] & 24 < SNR < 28, \\ [1 & 0 & 0 & 0] & 20 < SNR < 24 \\ [0 & 0 & 0 & 0] & SNR < 20 \end{cases} \quad (6-11)$$

where a 1 corresponds to voiced and 0 corresponds to unvoiced, going left to right the voicing V_{B_L} vector goes from band 1 to band 4, and the number of harmonics given in specific band is determined by equation 6-10.

6.1.4 Spectrum

As with EMBE, the goal of a spectral model for a harmonic based vocoder is to accurately represent the harmonic amplitudes for voiced speech, and to fit the spectrum in an average sense for unvoiced speech. Harmonic coders usually employ some direct form of quantization of the harmonic amplitudes to achieve this. While this results in a highly accurate representation of the spectrum, the number of bits required precludes its use for low and mid rate coding, depending on the frame rate. This is overcome with the use of a parametric model for the spectrum, such as Linear Prediction. Similar to EMBE, the SMABS vocoder represents the spectrum using a spline enhanced, linear predictive (LP) model for voiced speech, and a traditional LP model for unvoiced speech.

In this vocoder an LP model order of 18 is selected to model the spline envelope of the magnitude spectrum. The LP model coefficients are computed using a frequency domain approach, rather than the traditional time-domain approach. This allows the manipulation of the spectrum prior to model computation to enhance perceptually

important areas. This concept is the same as presented in Chapter 3 for the EMBE 2,400 bit per second speech coder. The main difference is that the SMABS vocoder is not using the warping function.

In the case that the current frame of speech is declared entirely unvoiced, the LP coefficients are computed using equations 3-57, 3-58, and 3-59 with no spectral amplitude compression or spline fit to the spectral amplitudes. If any band is declared voiced then the spectral amplitudes are compressed using a logarithm function and then a cubic spline is fit to the compressed spectral amplitudes as described by equations 3-60 through 3-69. An LP model is then fit to the spline envelope using equations 3-73 through 3-75.

Once the LP model coefficients have been calculated, they are converted to an alternate representation, known as line spectral pairs (LSP's). Line spectral pairs are known to exhibit superior quantization properties when compared to predictor coefficients. The LSP's are obtained by decomposing the impulse response of the LP analysis filter into difference and sum filters. These operations are shown in equations 3-76 – 3-78.

6.1.5 Gain

The gain value from the analysis-by-synthesis loop as presented in Chapter 5 for the time-domain analysis-by-synthesis method is no longer a valid gain term outside the loop since the phase information is not being transmitted. Thus it is sufficient to compute a gain for the LP model by calculating the ratio of the energy of the original spectrum and LP model spectrum, as is typically done.

This concept is slightly modified for the SMABS vocoder. For the voiced frames the gain is found by computing the ratio of the energy of the LP model spectrum and the energy of the cubic spline envelope. The gain for unvoiced frames is found in typical fashion by computing the ratio of the energy of the original spectrum and LP model spectrum.

6.2 Quantizer

Once the model parameters are calculated, they are quantized to approximately 8,000 bps for transmission. At this bit rate and update rate, 242 bits are available to represent all the parameters in each 30 ms interval. The gain, pitch, and voicing decisions are coded using simple scalar quantization, while the spectral model is coded using vector quantization of the line spectral pairs. Figure 6-2 summarizes the bit allocation and sub/superframe update scheme for each parameter, where there are two subframes and two superframes in each analysis frame.

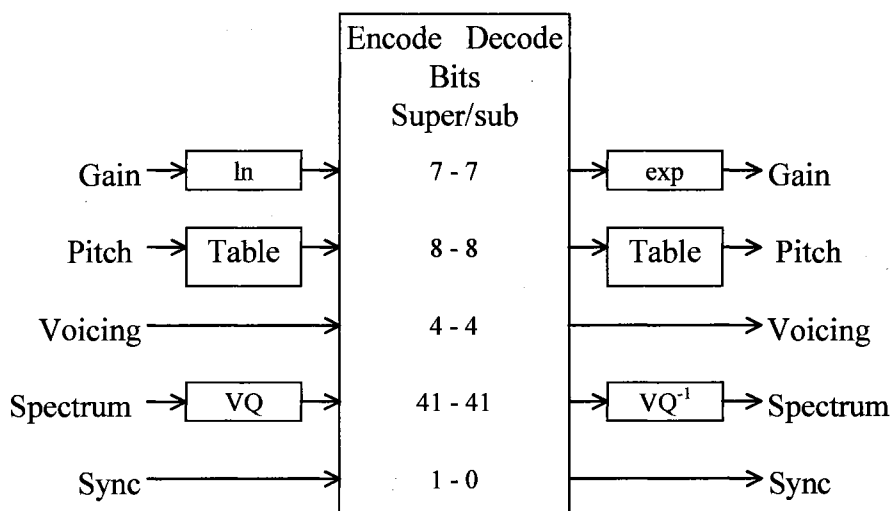


Figure 6-2. Block Diagram of Quantization and Coding

The voicing decisions are quantized as 4 bits with each bit corresponding to the voicing decision for the respective frequency band. As stated previously, the pitch is quantized linearly in samples from the range of 20 samples to 114 samples using 8 bits per subframe. The gain G is logarithmically scalar quantized as is done in EMBE following equations 3-90 and 3-91. The main difference is that there are 7 bits available for each subframe.

Prior to quantization, the LP coefficients are converted to line spectral pairs (LSP's). As stated, LSP's have superior transmission and quantization properties over traditional LP coefficients. A vector quantization (VQ) approach is used for coding the 18th order LSP model. One potential implementation is a 41-bit, 4 way split VQ codebook. The 4 way split is broken down to 11, 10, 10, and 10 bits respectively. The VQ codebooks are searched for each target LSP by minimizing the squared distance between the original LSP's and the target codebook vector. The split codebooks reduce the computational complexity by allowing each codebook to represent only a small segment of the LSP spectrum and reducing the amount of memory necessary to store the VQ tables.

6.3 Synthesizer

6.3.0 Introduction

In the synthesizer, each parameter vector is recovered by reversing the encoding procedure applied in the analyzer. The vocal tract spectrum, represented as LSPs, is converted back to the coefficients of an LP model. Similar to EMBE a sinusoidal model

is used to generate synthetic data on a frame-by-frame basis. In contrast to the method of reconstruction used in EMBE, a single method of reconstruction is used for both the voiced and unvoiced bands. A block diagram of the SMABS synthesizer is given in Figure 6-3.

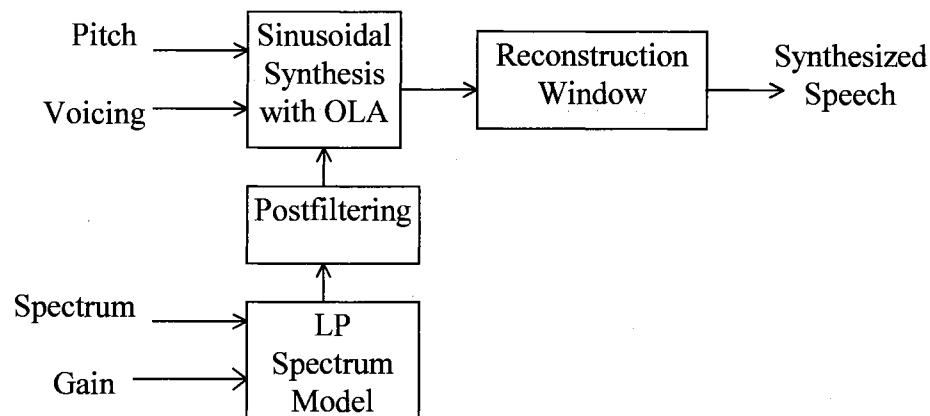


Figure 6-3. Block Diagram of SMABS Synthesizer

If the current band is declared voiced or unvoiced, a bank of sinusoidal oscillators is used to generate a periodic or aperiodic signal corresponding to each harmonic in the band. These harmonics are then scaled by the appropriate harmonic amplitudes, rotated by the appropriate phases, and summed and overlapped with the previous frame(s) producing a signal estimate for the current frame.

It is also possible that the gain varies substantially from one frame to the next, so some amplitude smoothing is needed. This is accomplished by overlapping the reconstructed speech frames using a Triangular reconstruction window. The window is designed so that the sum of the overlapped windows is unity.

The sections following, Spectral Filtering and Synthesis, describe the method used to generate synthetic speech using the sinusoidal model.

6.3.1 Spectral Filtering

This stage of the synthesizer refers to the processing of the spectral model transmitted by the analyzer. While this is presented as a prefiltering stage in the synthesizer it is most often referred to as postfiltering. Postfiltering is generally needed to aid in masking noise induced by the reconstruction process when no phase information is available, as is the case for this vocoder. The postfilter presented here is a slightly modified form of the filter presented in Chapter 3 for EMBE.

First, the spectral amplitudes are obtained by computing the LP spectrum and subsampling at the specified pitch and scaling by the respective gain. This postfilter is based on a design presented by [24] for sinusoidal based vocoders and is a variant of the postfilter presented in [22]. The amplitudes are weighted as given by

$$\tilde{S}_l = \alpha(\mu \hat{S}_l) \quad (6-12)$$

where \tilde{S}_l represents the postfiltered spectral amplitude and \hat{S}_l represents the spectral amplitude estimated from the LP model. The weighting is determined by μ , which is given by

$$\mu = \begin{cases} 1.2 & W_l > 1.2 \\ 0.5 & W_l < 0.5, \\ W_l & \text{otw} \end{cases} \quad (6-13)$$

where W_l is the weighting function given by

$$W_l = \hat{S}_l^\gamma \left[\frac{L(R_0^2 - 2R_0R_1 \cos(l\omega_o) + R_1^2)}{R_0(R_0^2 - R_1^2)} \right]^{\frac{\gamma}{2}}, \quad (6-14)$$

L represents the number of harmonics in the current frame and γ falls in the range between 0 and 1, where $\gamma = 0.5$ for the SMABS vocoder. The thresholds for W_l were determined experimentally in [24]. The R 's represent the energy and the first correlation coefficient and both are given by

$$R_0 = \sum_{l=1}^L \hat{S}_l^2 \quad (6-15)$$

and

$$R_1 = \sum_{l=1}^L \hat{S}_l^2 \cos(l\omega_o). \quad (6-16)$$

The last term to define is α . This coefficient is used to normalize the energy so that the original spectral amplitude estimates and the postfiltered spectral amplitudes have the same energy. This coefficient is given by

$$\alpha = \left[\frac{R_0}{\sum_{l=0}^L \hat{S}_l^2} \right]^{\frac{1}{2}} \quad (6-17)$$

6.3.2 Synthesis

The synthetic speech is generated on a frame-by-frame basis by decoding the transmitted parameters and applying them to the sinusoidal model. The speech is reconstructed using overlap addition. This method is chosen over that of EMBE because a single method is used to generate the voiced speech and the unvoiced speech.

This approach is detailed in Chapter 5, section 3. The speech for the current frame is computed using equation 5-13 and is then overlapped and added based on equation 5-16, where the reconstruction window is triangular shaped and satisfies the requirements of equation 5-17.

6.3.3 Reconstructed Output

The window used in the reconstruction process is an overlapping triangular window. This is the same window used in the analysis-by-synthesis loop defined in equation 6-9. The current frame produces 240 points of speech reconstructed using the sinusoidal model. The output is updated every 60 points which results in an overlap of four frames.

6.4 Conclusion

This chapter introduced a new analysis-by-synthesis vocoder known as Sinusoidal Model Analysis-By-Synthesis (SMABS). This vocoder used the time-domain analysis-by-synthesis procedure developed in Chapter 5 to estimate the model parameters for the sinusoidal model of reconstruction using OLA.

The synthetic speech generated using this vocoder is deemed to be of high quality based on informal listening tests. The synthetic speech produced is comparable to that of the frequency-domain analysis-by-synthesis approach but lower quality than that of the time-domain analysis-by-synthesis approach. The reason for this is the lack of phase information available in the receiver.

The computational complexity is still an issue. The frequency-domain analysis-by-synthesis approach was too computationally complex to be considered as a useful solution so the time-domain analysis-by-synthesis approach was selected for implementation. The computational complexity is contained in the number of sub-sampling periods that are searched. In the frequency-domain approach, the integer sub-sampling periods are searched and then a refinement stage is used to find the best sub-sampling period and its corresponding spectral amplitudes. The time-domain method uses much shorter correlations so the number of sub-sampling periods searched is increased. Even with this increase in the number of sub-sampling periods searched, the computational complexity of the time-domain approach is less than that of the frequency-domain approach.

7 CONCLUSION

7.0 Introduction

The sinusoidal model is chosen as the topic of this dissertation because sinusoidal based vocoders have been shown to be able to produce high quality speech at low bit rates. The main disadvantage of using the sinusoidal model in developing low bit rate vocoders is the high dependence on the parameter estimation, especially the pitch. The goal of this chapter is to apply the technique of analysis-by-synthesis to the problem of parameter estimation for sinusoidal based vocoders.

Assuming a frame of speech is modeled accurately using the sinusoidal model, as presented in Chapter 4, then a technique is needed to determine the appropriate set of amplitudes, frequencies, and phases (the model parameters) used to represent a frame of speech. The DFT is utilized in the analyzer for extracting the amplitudes, frequencies, and phases for the sinusoidal synthesis procedure. Determining these parameters is accomplished by developing an analysis-by-synthesis technique to improve the parameter estimation for sinusoidal based vocoders. Two novel analysis-by-synthesis methods are presented in Chapter 5. The first approach is developed in the frequency-domain and the second is developed in the time-domain.

The main difference between the methods developed in this dissertation and typical LP analysis-by-synthesis systems is the method used to perform the reconstruction. In this dissertation a sinusoidal synthesis procedure is included in the

analysis loop to determine the appropriate model parameters for the sinusoidal model. The main advantage for including the synthesis method in the analysis is to aid in determining the appropriate model parameters for a given reconstruction method. This leads to a closed-loop analysis-by-synthesis procedure for determining the sinusoidal model parameters. By using a closed-loop approach, the parameters of the model are varied in a systematic way to produce a set of parameters that produce a synthetic signal, which matches the original signal in a minimum mean-squared error sense.

7.1 Frequency-Domain Analysis-By-Synthesis

A frequency-domain analysis-by-synthesis method for determining the model parameters for a sinusoidal model was developed in this dissertation. The analysis and synthesis techniques described in sections 5.2 and 5.3 are combined to form a closed-loop analysis-by-synthesis procedure to estimate the model parameters for a sinusoidal model using a minimum mean-squared error. This approach assumes that no phase information is available, so the phase must be synthesized according to the methods presented in Chapter 5 and in Appendix A1.

A mathematical development that determines the optimum sub-sampling period and its corresponding spectral amplitudes is presented in Chapter 5. This is accomplished by first determining a set of spectral amplitudes by sub-sampling the original magnitude spectrum. The phase is then synthesized from the spectral envelope corresponding to the spectral amplitudes. Then based on the sub-sampling period and the corresponding spectral amplitudes a set of voicing decisions are determined using the MBE analysis model (multiple voicing decisions). The sub-sampling period, spectral amplitudes,

voicing decisions, and synthetic phase are applied to the sinusoidal model to produce an estimate for the original magnitude spectrum. The minimum mean-squared error between the original magnitude spectrum and the synthetic magnitude spectrum is found using a two-stage process. First an optimum gain term is computed. Then a match score between the target magnitude spectrum and the synthetic magnitude spectrum is computed based on the optimum gain and the parameters corresponding to the optimum gain. The parameters that correspond to the minimum mean-squared error are selected to represent the current analysis frame.

Since the assumption is that no phase information is available in the analyzer it must be generated in the synthesizer. The method for generating the synthetic phase is discussed in section 5.2 and is presented in greater detail in Appendix A1.

This frequency-domain analysis-by-synthesis approach was found to be sufficient to determine the model parameters for the sinusoidal model using OLA for reconstruction. The synthetic speech generated using the frequency-domain analysis-by-synthesis method using the sinusoidal model with OLA was deemed to be of high quality from informal listening tests.

The main drawback with the frequency-domain analysis-by-synthesis is the computational complexity that results from having to use a DFT of length 16,384. The optimum gain and match scores are found by computing the correlation between the synthetic magnitude spectrum and the target magnitude spectrum. A DFT length equal to 16,384 results in the correlation of sequences of length 8,192. This is done for each sub-sampling candidate in every analysis frame.

7.2 Time-Domain Analysis-By-Synthesis

A time-domain analysis-by-synthesis method for determining the model parameters for a sinusoidal model was also developed in this dissertation. Again the analysis and synthesis techniques described in sections 5.2 and 5.3 are combined to form a closed-loop analysis-by-synthesis procedure to estimate the model parameters for a sinusoidal model using a minimum mean-squared error. In contrast to the frequency-domain analysis-by-synthesis this approach assumes that phase information is available, thus time alignment is maintained.

The mathematical development that determines the optimum sub-sampling period and its corresponding spectral amplitudes is presented in Chapter 5. This is accomplished by first determining a set of spectral amplitudes and phases by sub-sampling the original magnitude spectrum and phase response. The sub-sampling period, spectral amplitudes, and phases are applied to the sinusoidal model to produce an estimate for the original time-domain signal. The minimum mean-squared error between the original time-domain signal and the synthetic time-domain signal is found using a two-stage process, similar to the frequency-domain analysis-by-synthesis. First an optimum gain term is computed. Then a match score is computed given the optimum gain and the parameters corresponding to the optimum gain. The parameters that correspond to the minimum mean-squared error are selected to represent the current analysis frame.

This time-domain analysis-by-synthesis approach was found to be sufficient to determine the model parameters for the sinusoidal model using OLA for reconstruction. The synthetic speech generated using the time-domain analysis-by-synthesis method using the sinusoidal model with OLA was deemed to be of high quality through informal

listening tests. The quality of the time-domain analysis-by-synthesis method is much higher than that of the frequency-domain analysis-by-synthesis method. This result was attributed to the fact that the phase information is available, instead of the less reliable voicing decisions. The time-domain approach is also very robust in a number of different environments such as quiet and office noise. The addition of the phase information also masked any potential problems with pitch halving or doubling.

The main drawback with the time-domain analysis-by-synthesis is that the phase information must be transmitted. The inclusion of the phase information forces the vocoder to operate at a much higher bit rate as compared to the frequency-domain analysis-by-synthesis approach.

7.3 Sinusoidal Model Analysis-By-Synthesis Vocoder

The frequency-domain analysis-by-synthesis method of parameter estimation is targeted at low bit rates but has high computational complexity, while the time-domain analysis-by-synthesis method of parameter estimation is targeted at higher bit rates with lower computational complexity. Therefore a combination of the two approaches is used to develop a new vocoder targeted for 8,000 bps.

The time-domain analysis-by-synthesis method is used to estimate the sinusoidal model parameters for this vocoder. The analysis and synthesis techniques described in sections 5.2 and 5.3 are combined with the time-domain analysis-by-synthesis method to form a closed-loop procedure to estimate the model parameters for a sinusoidal model using a minimum mean-squared error. This approach assumes that phase information is available for the analysis so that time alignment is maintained. This phase information is

not transmitted to the receiver in order to reduce the overall bit rate. The phase is synthesized in the receiver using the method presented in the frequency-domain analysis-by-synthesis method of parameter estimation.

This sinusoidal model analysis-by-synthesis (SMABS) vocoder was found to be sufficient to determine the model parameters for the sinusoidal model using OLA for reconstruction. The synthetic speech generated using this approach was deemed to be of high quality through informal listening tests. The quality of the time-domain analysis-by-synthesis method is similar to that of the frequency-domain analysis-by-synthesis but with lower computational complexity.

The main disadvantage of this vocoder was in the estimation of the voicing decisions, an unexpected problem.

7.4 Future Research

7.4.0 Introduction

The following paragraphs outline and discuss potential topics of future research for the two analysis-by-synthesis methods developed in this dissertation. The topic of computational complexity is discussed for both methods while the trade-off between “no phase” and phase is discussed in the case of the time-domain method.

7.4.1 Frequency-Domain Analysis-By-Synthesis

The frequency-domain analysis-by-synthesis method of parameter estimation is sufficient for determining the parameters of a sinusoidal model using overlap addition. The main disadvantage of this method is the computational complexity that results from

the long DFTs that are computed, which are necessary to obtain sufficient frequency resolution.

In the simulation a 16,384 point DFT is used to compute the magnitude spectrum of the original input signal and is used to compute the synthetic magnitude spectrum for each sub-sampling period candidate. Then to determine the best sub-sampling period, the correlation between the target magnitude spectrum and the synthetic magnitude spectrum is computed over half of the spectrum, in this case a correlation length of 8,192. If only the integer sub-sampling periods are considered this results in 95 DFTs of length 16,384 and 94 correlations of length 8,192 for each frame. A topic of future research for reducing the complexity is to develop a harmonic based DFT, which only compute the desired values (i.e., the harmonic values).

This seems reasonable since the only values used from the magnitude spectrum are determined by the sub-sampling period, which has a maximum number of 56 sample points at a sub-sampling period of 114 samples. In the worst case scenario, sub-sampling period equal to 114, only 56 of the 8,192 values are necessary. By developing a harmonic based DFT only the values necessary are computed. A M point DFT takes $(M/2)\log_2(M)$ number of complex multiplications and $M\log_2(M)$ number of complex additions. If $M = 16,384$ there are 114,688 complex multiplications and 229,376 complex additions are necessary. The problem lies in the fact the analysis frame contains only 240 points of real data, the rest of the values are zero as a result of the zero padding. If the DFT values can be computed in a harmonic fashion and all the zero multiplications are avoided, the equivalent would be a DFT of maximum length 56 (64 if rounded to a power of 2). The number of complex multiplications and additions necessary to compute

a $M = 64$ point DFT are 192 and 384, respectively. This results in considerable savings in the amount of computation necessary to perform the analysis-by-synthesis making it a more viable alternative.

The idea presented here is not meant to be an exact solution but is presented so as to promote thought in the area of determining the spectral amplitude and phases for sinusoidal based vocoders.

7.4.2 Time-Domain Analysis-By-Synthesis

The time-domain analysis-by-synthesis method of parameter estimation is sufficient for determining the parameters of a sinusoidal model using overlap addition. This approach has the advantage over other methods of not estimating any voicing decisions. This is reasonable since the phase information is being transmitted. The disadvantage with this approach is the higher bit rate, 16,000 bps. While this does code speech sampled at 8,000 samples per second using 2 bits per sample, a more useful application is in the area of mid and low bit rate coding, 8,000 bps and below.

The sub-sampling period and gain can be coded with a minimal number of bits and methods exist to code the spectral amplitudes in an efficient manner. So the problem here is to find a method to code the phase information efficiently. This would provide a big improvement in the quality, intelligibility, and robustness of sinusoidal based vocoders.

BIBLIOGRAPHY

- [1] A. Gersho, "Advances in Speech and Audio Compression," *Proceedings of the IEEE*, Vol. 82, No. 6, June 1994.
- [2] L. Rabiner, "Applications of Voice Processing to Telecommunications," *Proceedings of the IEEE*, Vol. 82, No. 2, February 1994.
- [3] R. McAulay and T. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, August 1986.
- [4] D. Griffin and J. Lim, "MultiBand Excitation Vocoder," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. ASSP-36, No. 8, August 1988.
- [5] D. Griffin, "MultiBand Excitation Vocoder," MIT, USA, 1987, Ph.D.
- [6] W. Andrews, "Design of a High Quality 2400 Bit Per Second Enhanced MultiBand Excitation Vocoder", M.S. Thesis, E.C.E.N. Department, Oklahoma State University, 1994.
- [7] K. Teague, B. Leach, and W. Andrews, "Development of a High-Quality MBE based Vocoder for Implementation at 2400 bps," *Proceedings IEEE Wichita Conference on Communications, Networking, and Signal Processing*, April 1994.
- [8] B. Atal, "Predictive Coding of Speech at Low Bit Rates," *IEEE Transactions on Communications*, pp. 600-614, April 1982.
- [9] A. Spanias, "Speech Coding: A Tutorial Review," *Proceedings of the IEEE*, Vol. 82, No. 10, October 1994.
- [10] J. Deller, J. Proakis, and J. Hansen, "*Discrete-Time Processing of Speech Signals*," Macmillan Publishing Company, New York, 1993.
- [11] D. O'Shaughnessy, "*Speech Communication: Human and Machine*," Addison-Wesley, New York, 1987.
- [12] L. Rabiner and R. Schafer, "*Digital Processing of Speech Signals*," Prentice Hall, New Jersey, 1978.
- [13] H. Edwards, "*Applied Phonetics: The Sounds of American English*," Singular Publishing Group, San Diego, 1992.
- [14] F. Owens, "*Signal Processing of Speech*," McGraw Hill, New York, 1993.

- [15] C. McElroy, "Wideband Speech Coding," Department of Electrical and Electronic Engineering, University College, Dublin, Ph.D., 1994.
- [16] A. Gersho and R. Gray, "*Vector Quantization and Signal Compression*," Kluwer Academic Publishers, 1992
- [17] R. Gray, "Vector Quantization," *IEEE Acoustics, Speech, and Signal Processing Magazine*, April 1994.
- [18] J. Makhoul, S. Roucos, and H. Gish, "Vector Quantization in Speech Coding," *Proceedings of the IEEE*, Vol. 72, No. 11, November 1985.
- [19] Inmarsat Satellite Communications Services, "Inmarsat-M System Definition, Issue 3.0-Module 1: System Description," November 1991.
- [20] J. Campbell *et al*, "The proposed Federal Standard 1016 4,800 bps Voice Coder," *Speech Technology*, pp. 58-64, April 1990.
- [21] J. Campbell, T. Tremain, and V. Welch, "The Federal Standard 1016 4,800 bps CELP Voice Coder," *Digital Signal Processing*, Academic Press, Vol. 1, pp. 145-155, 1991.
- [22] Federal Standard 1016, "Telecommunications: Analog to Digital Conversion of Radio Voice by 4,800 bit per second Code Excited Linear Prediction (CELP)," National Communications System, Office of Technology and Standards, Washington, DC 20305-2010, 14 February 1991.
- [23] D. Kemp, R. Sueda, and T. Tremain, "Evaluation of 4,800 bps Voice Coders," *Proceedings of the Military and Government Speech Technology*, November 1989.
- [24] R. McAulay and T. Quatieri, "Sinusoidal Coding," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), pp. 121-173, Amsterdam, The Netherlands, Elsevier Science, 1995.
- [25] D. Griffin, and J. Lim, "A New Model Based Speech Analysis/Synthesis System," *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1985.
- [26] A. Kondo, *Digital Speech, Coding for Low Bit Rate Communications Systems*, pp. 79-115, West Sussex, England, John Wiley & Sons, 1994.
- [27] APCO, "NASTD Federal Project 25 Vocoder: Version 1.0," December 1992.

- [28] K. Teague, W. Andrews, and B. Walls, "Enhanced Modeling of Discrete Spectral Amplitudes," *IEEE Speech Coding Workshop*, September 1997.
- [29] K. Teague, and W. Andrews, "Enhanced Spectral Modeling for MBE Speech Coders," *IEEE 31st Asilomar Conference*, November 1997.
- [30] J. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-25, pp. 367-377, Dec. 1972.
- [31] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, Vol. 63, No. 4, April 1975.
- [32] J. Makhoul, "Spectral Linear Prediction: Properties and Applications," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-23, No. 3, pp. 283-296, June 1975.
- [33] R. Sedgewick, *Algorithms in C*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1990.
- [34] G. Kang, and L. Fransen, "Application of Line Spectral Pairs to Low Bit Rate Speech Encoders," *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1985.
- [35] N. Jayant and P. Noll, "Digital Coding of Waveforms," Prentice-Hall, New Jersey, 1984.
- [36] K. Paliwal and B. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame," *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No. 1, pp. 3-14, January 1993.
- [37] A. Oppenheim and R. Schaffer, "Discrete-Time Signal Processing," Prentice Hall, New Jersey, 1989.
- [38] R. Zeimer and W. Tranter, "Principles of Communications: Systems, Modulation, and Noise," Houghton Mifflin, Boston, 1990.
- [39] R. Ramachandran and W. Mammone, "Modern Methods of Speech Processing," Kluwer Academic Publishers, Boston, 1995.
- [40] R. Churchill and J. Brown, "Complex Variables and Applications," McGraw Hill, New York, 1990.
- [41] F. Minifie, T. Hixon, and F. Williams, "Normal Aspects of Speech, Hearing, and Language," Prentice Hall, New Jersey, 1973.

APPENDIX

A.1 Synthetic Phase

The generation of high quality speech using the sinusoidal model is dependent on an accurate phase track at least for harmonic amplitudes, which are declared voiced. It is well known that during voiced speech the production of speech begins with a sequence of excitation pitch pulses that represent the closure of the glottis at a rate determined by the pitch frequency. This suggests that a linear phase model is sufficient for modeling the phase. It is noted that in general the harmonic amplitudes may not be harmonically related to the pitch. The altering of the spectral amplitudes and corresponding phase is modeled by the transfer function $H_s(\omega)$, which is defined to be a minimum phase system [24]. This transfer function is made up of the composite of two transfer functions written as

$$H_s(\omega) = |H_s(\omega)| e^{j\Phi_s(\omega)}, \quad (\text{A.1-1})$$

where this composite function is referred to as the system function and $|H_s(\omega)|$ represents the magnitude and $\Phi_s(\omega)$ represents the system phase and $j = \sqrt{-1}$. The harmonic phase model is now defined in terms of the linear portion plus the system phase given by

$$\theta_l = \phi_0 l + \Phi_s(\omega), \quad (\text{A.1-2})$$

where ϕ_0 represents the starting phase corresponding to the fundamental frequency and $1 \leq l \leq L$.

The starting phase is determined by computing the integral of the instantaneous frequency given by

$$\phi_0 = \phi_0^{k-1} + \int_T \omega_0(\sigma) d\sigma, \quad (\text{A.1-3})$$

where ϕ_0^{k-1} represents the starting phase from the previous frame, $\omega_0(\sigma)$ represents the time-varying property of the pitch frequency, and T is defined to be the length of the synthesis frame. The pitch frequency is defined to be a linearly varying function from frame-to-frame and is written in terms of the previous frame's pitch frequency and the current frame's pitch frequency. This function represents an average pitch and is written as

$$\omega_0(t) = \omega_0^{k-1} + \frac{\omega_0^k - \omega_0^{k-1}}{T} t. \quad (\text{A.1-4})$$

The substitution of this function into the integral equation above results in a representation for the starting phase in the current frame. This is given by

$$\phi_0 = \phi_0^{k-1} + \frac{(\omega_0^{k-1} + \omega_0^k)T}{2}. \quad (\text{A.1-5})$$

The previous equations define the computation of linear phase portion of the phase model. The following equations define the computation of the system phase portion. As stated above the system function is defined to be a minimum phase system. If this assumption is true then the system function is can be expressed in terms of the cepstral coefficients.

We start by computing the complex logarithm of the function $H_s(\omega)$ and then compute the inverse Fourier transform to obtain the cepstral coefficients. The complex logarithm [40] is given as

$$\log(H_s(\omega)) = \ln(|H_s(\omega)|) + j\Phi_s(\omega). \quad (\text{A.1-6})$$

In computing the inverse Fourier transform the system phase term is ignored since only the magnitude function $|H_s(\omega)|$ is available. The inverse Fourier transform is written as

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln(|H_s(\omega)|) e^{j\omega n} d\omega. \quad (\text{A.1-7})$$

Since the magnitude function $|H_s(\omega)|$ has the property of being an even function then the inverse Fourier transform is written solely in terms of a cosine, as given by

$$c_n = \frac{1}{\pi} \int_0^{\pi} \ln(|H_s(\omega)|) \cos(\omega n) d\omega. \quad (\text{A.1-8})$$

The system phase is now computed using the Fourier transform. This is defined by

$$H_s(\omega) = \sum_n c_n e^{-j\omega n}. \quad (\text{A.1-9})$$

This equation is now rewritten by substituting the magnitude and phase functions for $H_s(\omega)$ and rewriting the complex exponential in terms of cosines and sines. The Fourier transform is now given by

$$\ln|H_s(\omega)| + j\Phi_s(\omega) = \sum_n c_n \cos(\omega n) - jc_n \sin(\omega n). \quad (\text{A.1-10})$$

This equation is now split into the real and imaginary parts, shown below, resulting in the following equations for the logarithm magnitude function and system phase function and exploiting the symmetry properties.

$$\ln|H_s(\omega)| = 2 \sum_n c_n \cos(\omega n) \quad n = 0, 1, \dots, N \quad (\text{A.1-11})$$

$$\Phi_s(\omega) = -2 \sum_n c_n \sin(\omega n) \quad n = 1, 2, \dots, N \quad (\text{A.1-12})$$

For a speech signal sampled at 8,000 samples per second a value of N greater than or equal to 44 has been determined to be sufficient for computing the system phase. The harmonic phase model is now defined in terms of the linear model and the system phase given by

$$\theta_l = l \left(\phi_0^{k-1} + \frac{(\omega_0^{k-1} + \omega_0^k)T}{2} \right) - 2 \sum_{n=1}^N c_n \sin(\omega n). \quad (\text{A.1-13})$$

A.2 Postfilter

The sinusoidal model of speech reconstruction has been shown to produce high quality and intelligible speech. Although the reconstructed speech, typically has a muffled quality. A number of efforts have been directed towards correcting for this muffled quality which is result of the sinusoidal model analysis [24]. The biggest effort to get rid of this coder noise is in the area of auditory masking or postfiltering of the spectral envelope. The approach used in this dissertation is a variant of the FS1016 CELP postfilter. Since the analysis and synthesis are performed in the frequency-domain, it makes sense that the postfiltering should occur in the same domain.

The idea is to design the postfilter such that the dynamic range between the formant peaks and formant nulls of the spectral envelope is increased. This is accomplished using a filter that performs a spectral tilt of the spectral envelope as given by

$$W(\omega) = \frac{|H_s(\omega)|}{|T(\omega)|}, \quad (\text{A.2-1})$$

where $T(\omega)$ represents the tilting function, $H_s(\omega)$ is spectral envelope, and $F(\omega)$ is the spectrally flattened version of $H_s(\omega)$.

In order to perform compression of $F(\omega)$ then it must be normalized to have unity gain. If this is true then the compression function is given by

$$C(\omega) = [F(\omega)]^\alpha, \quad (\text{A.2-2})$$

where $0 \leq \alpha \leq 1$. The spectral tilting function can be defined by a simple first order all pole model written as

$$T(\omega) = \frac{\kappa}{1 - \mu e^{-j\omega}}, \quad (\text{A.2-3})$$

where μ is found using LPC analysis techniques on the synthetic speech waveform and κ represents the gain of the filter. This coefficient represents the first prediction coefficient, which is found from the ratio of the energy R_0 and the first correlation coefficient R_1 . It can be shown that these correlation coefficients can be found in the frequency-domain and are given by

$$R_0 = \sum_{l=1}^L \hat{S}_l^2 \quad (\text{A.2-4})$$

and

$$R_1 = \sum_{l=1}^L \hat{S}_l^2 \cos(l\omega_0), \quad (\text{A.2-5})$$

where $\mu = R_0 / R_1$, L determined the number of harmonics in the current synthesis frame, and ω_0 is the pitch frequency for the current synthesis frame.

By making the appropriate substitutions the spectrally flattened spectral envelope is now written as

$$F(l\omega_0) = \hat{S}_l \left[\frac{(1 + \mu^2) - 2\mu \cos(l\omega_0)}{\kappa^2} \right]^{\frac{1}{2}}. \quad (\text{A.2-6})$$

Since $F(\omega)$ is normalized to have unity gain, κ is chosen such that the average power of the spectrally flattened amplitudes is unity and is given by

$$\frac{1}{L} \sum_{l=1}^L F(l\omega_0)^2 = 1 = \frac{1}{L\kappa^2} \left[(1 + \mu^2) \sum_{l=1}^L \hat{S}_l^2 - 2\mu \sum_{l=1}^L \hat{S}_l^2 \cos(l\omega_0) \right]. \quad (\text{A.2-7})$$

This equation is now solved for the gain μ and making the appropriate substitutions results in

$$\kappa^2 = \frac{1}{L} \left[(1 + \mu^2) R_0 - 2\mu R_1 \right]. \quad (\text{A.2-8})$$

By substituting the gain term into the equation for the spectrally flattened spectral envelope the result is a given by

$$F(l\omega_0) = \hat{S}_l \left[\frac{L[R_0^2 + R_1^2 - 2R_0 R_1 \cos(l\omega_0)]}{R_0(R_0^2 - R_1^2)} \right]^{\frac{1}{2}}. \quad (\text{A.2-9})$$

The values of the postfilter are given by raising the spectrally flattened spectral envelope to the appropriate power as defined previously. The postfilter is now written as

$$C_l = \hat{S}_l^\alpha \left[\frac{L[R_0^2 + R_1^2 - 2R_0R_1 \cos(l\omega_0)]}{R_0(R_0^2 - R_1^2)} \right]^{\frac{\alpha}{2}}. \quad (\text{A.2-10})$$

After postfiltering the spectral amplitude, the energy in the postfiltered amplitudes is no longer equal to the original spectral amplitudes so it is necessary correct the energy level in the synthetic signal. This is accomplished by scaling the postfiltered amplitudes such that the energy is the same as before the postfiltering defined by

$$\sigma = \left[\frac{R_0}{\sum_{l=1}^L C_l^2} \right]^{\frac{1}{2}}. \quad (\text{A.2-11})$$

The postfilter described above is the one used in this dissertation. There are other postfilter forms available. One alternate form of postfilter is used in the EMBE 2,400 bps vocoder and is described in Chapter 3 of this dissertation.

2

VITA

Walter D. Andrews

Candidate for the Degree of Doctor of Philosophy

Thesis: SINUSOIDAL MODEL ANALYSIS-BY-SYNTHESIS FOR SPEECH CODING

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in Clovis, New Mexico, March 18, 1965, the son of Kay Parker and Walt Andrews Jr.; married to Ronda Andrews and have two children, Chuck and Caleb.

Education: Graduated from Sapulpa High School, Sapulpa, Oklahoma, in May, 1983; received an Associate of Applied Science Degree in Electrical-Electronics Technology from Oklahoma State University Technical Branch Okmulgee, Okmulgee, Oklahoma, in September, 1986; received Bachelor of Science Degree in Electrical and Computer Engineering from Oklahoma State University, Stillwater, Oklahoma, in December, 1993; received Master of Science Degree in Electrical and Computer Engineering from Oklahoma State University, Stillwater, Oklahoma, in December, 1994; completed requirements for the Doctor of Philosophy at Oklahoma State University, Stillwater, Oklahoma, in May, 1998.

Professional Experience: Engineering Intern, Los Alamos National Lab, Los Alamos, New Mexico, June, 1991 to August, 1991; Engineering Intern, REDA Pump, Bartlesville, Oklahoma, June, 1992, to August, 1992; Student Instructor, Department of Electrical and Computer Engineering, Oklahoma State University, Fall 1995 and Fall 1996; Research Assistant, Department of Electrical and Computer Engineering, Oklahoma State University, June, 1993 to present.

Professional Memberships: IEEE, Eta Kappa Nu, and Tau Beta Pi.